DOCUMENT RESUME

ED 068 491 TM 001 827

AUTHOR Light, Judy A.
TITLE The Development and Application of a Structured Procedure for the In-Context Evaluation of Instructional Materials.
BUREAU NO BR-5-0253
PUB DATE 72
NOTE 109p.; Submitted to the Graduate Faculty in the School of Education in partial fulfillment of the requirements for the degree of Masters of Arts, University of Pittsburgh

EDRS PRICE MF-$0.65 HC-$6.58
DESCRIPTORS Classroom Environment; *Classroom Materials; *Curriculum Evaluation; *Educational Improvement; *Evaluation Techniques; Grade 4; Instructional Materials; Learning Activities; Learning Motivation; Mathematics Curriculum; Systems Analysis; Test Construction

ABSTRACT

An attempt was made to develop and demonstrate the use of a clearly specified procedure for identifying specific causes of system failures as these are encountered in the in-classroom tryout of new lesson materials and their associated classroom management activities. Evaluation models were reviewed as well as the literature on curriculum evaluation procedures and designs. The study was conducted in two fourth grade classes using the Individually Prescribed Instruction mathematics curriculum. A specific evaluation technique was designed based on a quasi-experimental design for establishing cause and effect relationships. The four steps used in applying the model were defining the goals, plan, operation, and assessment. The results indicate that the procedures developed for this study were successful in (1) monitoring the total classroom environment by identifying those variables that interfere with learning, and (2) attempting to define and apply strong inference procedures for evaluating instructional materials. (JS)

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

THE DEVELOPMENT AND APPLICATION OF A STRUCTURED PROCEDURE FOR THE IN-CONTEXT EVALUATION OF INSTRUCTIONAL MATERIALS

Judy A. Light

B.A., Chatham College, 1967

TM 001 827

Submitted to the Graduate Faculty in the School of Education in partial fulfillment of the requirements for the degree of Masters of Arts

University of Pittsburgh

1972

ACKNOWLEDGEMENTS

The research reported herein was supported by the Learning Research and Development Center supported in part as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare. The opinions expressed do not necessarily reflect the position or policy of the Office of Education and no official endorsement should be inferred.

The writer wishes to express her deep appreciation to Mrs. Faye Mueller and Mrs. Joan Sherry for their willingness and cooperation in permitting the writer to conduct this study in their classes. Without their commitment and enthusiasm this study would never have been successful.

The writer also wishes to thank Dr. Larry Reynolds for his continual support and encouragement throughout this study.

TABLE OF CONTENTS

		Page
I.	INTRODUCTION	1
	A. Purpose of Study	1
	B. Review of Evaluation Models	4
	C. Review of Literature-- Curriculum Evaluation Procedures	7
	D. Review of Designs	12
II.	THE PROBLEM	19
III.	PLAN OF THE STUDY	23
	A. Instructional System Under Study	24
	B. Design	24
	C. Procedures	28
	D. Data Gathered	31
IV.	RESULTS	33
	A. Description of Instructional System	33
	B. The Identification of Inadequate Materials	41
	C. Possible Causes of Test Failure	44
	D. Construction of Small Inference Trees	47
	E. Comprehensive Inference Trees	65
	F. The Results of the Formative Evaluation of IPI Materials	80
V.	CONCLUSIONS AND SUMMARY	83
	A. Conclusions	83
	B. Summary	84

TABLE OF CONTENTS (continued)

Page

C. Suggestions for Additional Research. . . . 88

Appendix A 91

Bibliography 94

LIST OF TABLES

Page

Table 1 Plan for the In-Context Tryout of Instructional Materials 38

LIST OF FIGURES

		Page
Figure I	A Student's Responses to a Four-Item Test on Multiplication	48
Figure II	The Last Page Worked Before the Test in Figure I	50
Figure III	The Revised Page Added to the Materials.	52
Figure IV	Another Student Fails the Same Test	53
Figure V	The Revisions of the New Revised Page	56
Figure VI	A Student's Responses to a Five-Item Test on Numeration.	57
Figure VII	The Student's Responses to the Seven-Item Test of the Prerequisite Behaviors in Numeration	59
Figure VIII	A Student's Responses to a Nine-Item Test on Multiplication.	61
Figure IX	The Last Page Completed by the Student Before Taking the Test in Figure VIII	63
Figure X	Analysis of Failed CET Items	69
Figure XI	Analysis of Failed Posttest Items.	70
Figure XII	Analysis of Instructional Materials for Student Using Incorrect Rule	71
Figure XIII	Analysis of Instructional Materials for Improper Student Behavior	73
Figure XIV	Analysis of Instructional Materials for Unique Items	74

LIST OF FIGURES (continued)

		Page
Figure XV	Analysis of CET for Computational Posttest Errors	75
Figure XVI	Analysis of Systematic Posttest Errors.	76
Figure XVII	Analysis of Unsystematic Posttest Errors.	77
Figure XVIII	Analysis of Process Posttest Errors	78

CHAPTER I

INTRODUCTION

A. Purpose of Study

Evaluation, which may be defined as the collection and use of information to make decisions about a program, (Stufflebeam, 1970) can serve many roles in the design, implementation, and final assessment of an educational product. Cronbach (1963) and Hastings (1966) have discussed two major roles evaluation can serve: evaluation for decisions concerning adoption of a final product and evaluation for revisions of a product as it is being developed. Scriven (1967) has labeled these two types of evaluation as summative and formative evaluation. Until recently, more work has been done in summative evaluation, where the complete program is assessed, as opposed to formative evaluation where information is gathered and used as a basis for refinements in the program. The importance of the careful evaluation of materials during their construction has been widely recognized but few have attempted to outline specific guidelines for using formative evaluation in curriculum development.

This apparent lack of clearly specified procedures for using evaluation in curriculum development may arise

partly from the difficulties encountered in working in an on-going classroom, which is usually a necessity in carrying out formative evaluation. Chitayat (1970) suggests that it is difficult to identify all the variables in a classroom which can effect academic performance and to control for their direct and their interactive effects.

Frequently materials are "evaluated" in a pre-tryout session with several students in a non-school setting, then instituted on a large scale into the school, where only summative evaluation takes place. During the pre-tryout sessions, gross inadequacies associated with use in a regular classroom may not be detected.

In carrying out the in-context tryout of new lesson materials the developer must necessarily be concerned with all aspects of classroom operation that can affect pupil performance on the lessons. He must be concerned with how the lessons are used by teacher and pupil, with the degree to which specified procedures are followed, with pupil motivation, with the validity of testing procedures that are used, and with a number of other components of classroom operation. With any lesson it is assumed that it is used under certain conditions. The lesson can be given a meaningful tryout only if these conditions exist. In-classroom evaluation, then, must include the study of the extent to which the necessary conditions are present. When the person carrying out this type of evaluation

obtains information such as that pupils are not mastering a given skill that is covered in a part of the lesson, he cannot immediately assume that this part of the lesson is at fault. He must be concerned with other hypotheses that could explain this lack of mastery. Is the lesson being used properly? Is the criterion test a valid one? Was the pupil motivated to learn this skill? These and other alternative hypotheses must be investigated before a decision is made to revise the lesson.

It becomes apparent that a curriculum writer should no longer be concerned only with the lesson materials. He must also define other elements of the environment in which materials are to be used. He should define the behaviors of the teacher and the student, the work skills each pupil is to use, the information the teacher is to give each pupil and how the necessary skills can be taught.

The curriculum developer and the formative evaluator should not work independently of each other. The formative evaluator's role includes helping the curriculum developer attain the best environment for the use of the materials. His role should also include the observation of the total system in operation in order to offer suggestions for the modification of any components of the instructional system until a satisfactory environment and curriculum are obtained.

It is the purpose of this present study to develop, and demonstrate the use of a clearly specified procedure

for identifying specific causes of system failures as these are encountered in the in-classroom tryout of new lesson materials and their associated classroom management activities. The procedure will offer methods for identifying, controlling, and monitoring the factors which effect academic behavior in a classroom. It will deal with (1) the definition of management rules and ways to monitor their effectiveness, (2) the collection of objective and subjective data to discover weaknesses in both the materials and the classroom environment, and (3) the ways to systematically evaluate the effectiveness of all changes made in the environment. Although this procedure will be designed for evaluating individualized instruction materials, within certain limitations it should be useable with other curriculum programs.

B. Review of Evaluation Models

In order to organize the different roles of an evaluator into a workable relationship, several evaluation models have been proposed which present similar strategies for evaluating an educational program. Stake (1967) suggests a model for processing descriptive evaluation data. He states there are two main ways of using descriptive data to evaluate a program: finding contingencies among antecendents, transactions and outcomes, and finding congruence between intents and

observations. All programs have intended characteristics which must be compared to the observed characteristics of the program in operation. The observation of both the intended and observed characteristics of a program should be divided into three sub-categories: (1) antecedents, any condition which exists before the program, (2) transactions, interactions which occur during the program's operation, and (3) outcomes. The data is congruent if what is intended is what is actually observed. To establish logical contingencies, a logical connection between an event and purpose should exist. The logic of a contingency between intents is usually based on previous experience with similar observables rather than the direct observation. The contingency criterion between observed categories is based on empirical evidence rather than logical. The criterion is based on direct evidence of finding specific reasons for specific outcomes.

Lindvall and Cox (1969) offer a model for formative evaluation in Individually Prescribed Instruction (IPI). They outline four major steps in planning an education program: defining the goals, plan, operation, and assessment. The first requirement is that the goals of the program be well-defined and clearly stated. They should be observable, worthwhile, and attainable. Next a detailed plan of operation to achieve the suggested goals is necessary. The plan should be constructed to

insure successful attainment of the goals. The operation is the plan put into action. The main focus is on how successful the plan has been put into operation. Modifications to improve the implementation can be made. The final step is the assessment of the extent to which the goals have been met.

Stufflebeam's model (1970) for evaluation also uses categories similar to the IPI and Stake models. Stufflebeam suggests there are four types of data which can be collected and, therefore, four types of evaluation. Context evaluation aids in the selections of objectives by identifying problems which could prevent the attainment of needs and use of possible opportunities. Input evaluation, after analyzing alternative designs, decides on the best plan to achieve stated objectives. Process evaluation monitors the design. Product evaluation assesses the goals and reasons for the obtained results.

All three models appear to use different terminology to express the four major components necessary for a program's evaluation: a statement of objectives (intended outcomes, goals, context evaluation), methods for evaluating the obtained results (observed outcomes, assessment, product evaluation), a description of what the program should look like and a comparison of this with what the program is in operation (intended and observed transactions, plan and operation, input and process evaluation).

Although all three models imply the successive steps for evaluating a program as being a one-way progression, the use of formative evaluation implies the continual modifying of all steps at any time until a satisfactory final product is obtained.

C. Review of Literature--Curriculum Evaluation Procedures

The previous models outline the required components for developing and evaluating educational programs but they do not yield any information about practical specifics. Assessing the goals of a program should not only yield information about how well goals were met but should also include information about why they were not met. The causes behind failure are necessary in order to modify the goals and operation of the program to insure future success.

The practicalities of formative evaluation should include information on how to obtain and interpret data, how to identify and control variables which affect academic performance and how to use the data for refinement of the program. A model illustrating a complete evaluation procedure which includes defining these practicalities is needed.

In many educational programs the most obvious and accessible information is a set of test scores. Since poor test results could be a function of poor pupil behavior,

poor lesson materials, poor test structure, or poor teacher practices, they offer little help for evaluating materials. A model which controls these characteristics, analyzes test failures for the causes and corrects them, is necessary for formative evaluation.

Little work has been done in finding appropriate assessment procedures for curriculum development except in the area of programmed instruction. In these materials, student responses to each frame are individually evaluated for their effectiveness. Since evaluating programmed instruction involves evaluating the success of each frame, it seems appropriate to follow similar procedures in evaluating any curriculum materials.

Programmed instruction is based on Skinner's principle (Lumsdaine, 1964, p. 383) that any educational subject matter is an accumulation of behaviors which can be analyzed logically, and behaviorally into small successive steps. It is, therefore, logical to assume that an optimal sequence can be developed from analyzing the detailed records of each student's responses (Lumsdaine, 1964).

Two major factors relate to the evaluation of all instructional materials (Lumsdaine, 1965). First is the specification of the conditions under which the materials are to be used. This includes specifying how long the students worked, under what supervision, and with what

incentive. Secondly, influences from extraneous sources have to be controlled. The specification of the conditions and the control of extraneous sources are necessary requirements in evaluating any instructional materials. In order to isolate the reasons for given results, an attempt should be made to control all factors, other than the program itself, which could effect student performances.

The formative evaluation of programmed instructional materials does not include an experimental-control group design where each group randomly receives different mateials or related systems. Instead the procedures include rewriting materials until they are successful under certain specifications and controls within the instructional environment. The materials are evaluated on the basis of their effectiveness in teaching the objectives rather than by being compared with other programs.

Since the construction of completely successful instructional materials on the first attempt would be unusual, lesson writers advocate a pre-tryout stage to detect gross inadequacies in the materials (Nitko, 1968; Conrad, 1966; Markle, 1964; Gilbert, 1962). During a pre-tryout the lessons are used in an environment which is tightly controlled and removed from the classroom situation. Several programmers (Markle, 1964; Gilbert, 1962) suggest the lesson writer go through the materials

assuming the role of the student. It is during this time he can identify difficult or ambiguous items, poor directions, or poor logical progressions. After revising the materials the next step is to have several students, as similar to those in the final population as possible, use the materials. Suggestions for the number of students to use in pre-tryouts range from four to ten (Markle, 1964; Taber, Glaser and Schaefer, 1965).

Gilbert (1962) has isolated seven classes of behavior which should provide cues to defects in the materials. Type I behaviors are defects within the materials which rewriting would eliminate; (1) the student fails to complete a response, (2) the student makes an erroneous response due to irrelevant properties of the original stimulus, (3) the student makes an erroneous response which the lesson writer cannot account for, and (4) the student makes an erroneous response due to a competing response. A second class of errors, Type II, are not defects in the materials but are valuable in diagnosing difficulties. (1) The student hesitates or is perplexed by the materials, (2) the student is bored or tired, and (3) the student does not follow the proper sequence through the materials.

Gilbert also sets down rules for the observer with regard to intervention. He suggests that the observer refrain from giving any verbal instruction apart from that

specified by the program, but introduce changes if failure becomes certain and if it is necessary, question the student about his failure.

Although Gilbert (1962) is considering programmed instruction in a pre-tryout setting, attempting to use similar procedures for evaluating other types of instructional material in an in-context tryout could be valuable. Chitayat (1970) investigated three evaluation procedures for an in-context developmental tryout: (1) in the actual classroom with an observer present, (2) in a group where all students were working in the same unit and supervised by the evaluator, and (3) in a small group (four students) supervised and observed by the evaluator. More information about system variables was gained as the setting moved further away from that which the materials were intended.

Since Chitayat's (1970) and Gilbert's (1962) systems permit the evaluator to interact with individual students in a controlled setting, their procedures are not useful for evaluating instructional materials in the in-context setting. In-context evaluation must take place under conditions such that the regular instructional staff is in control of the classroom with no intervention by the evaluator. The purpose of the present study was to design and evaluate procedures for use in this latter type of setting.

D. Review of Designs

One of the major problems in doing research in curriculum development has been in selecting an appropriate design. True experimental designs have little applicability or feasibility in formative evaluation studies at this time. One of the requirements for a true experimental design is the random selection of groups (Campbell and Stanley, 1963) which can be considered equivalent on all crucial dimensions except for the treatment they receive with respect to the experimental variables. Although in formative evaluation studies random selection of students and teachers is possible, it would be difficult to maintain equivalent classroom conditions between groups. Since classroom conditions, such as teacher behavior and motivation, which effect academic performance, are heavily dependent upon the individual teacher's style, it appears impractical to assume that equivalent treatments could be maintained. Also since formative evaluation is concerned with answering questions about the quality of each component of the program, an experimental design comparing two groups would offer little information about the causes of failure which are necessary for the development of a program.

Since the use of true experimental designs appears at this time inappropriate for use with formative evaluation procedures, curriculum developers can seek other types of

designs to establish causal relationships similar to the "persuasive causal interpretations made possible by experiments involving randomization" (Campbell, 1963, p. 213).

Campbell and Stanley (1963) recognized that in certain natural settings the full control of the experimental variables cannot always be obtained. They have identified a group of "quasi-experimental designs" which can be used in situations where true experimental designs are not practical. These quasi-experimental designs can establish causal relationships under two conditions: that the interpretations made from the collected data must seem plausible, and other plausible rival hypotheses can be eliminated (Campbell, 1963). Campbell and Stanley (1963) have listed twelve threats to validity which form a list of probable rival hypotheses. Certain quasi-experimental designs control for some of these sources of invalidity. In other designs they form probable rival hypotheses which have to be considered as possible alternative interpretations.

Sidman (1960) suggests there are only two criteria, reliability and generality, which should be considered in accepting or rejecting data. Reliability can be established by repeating similar experiments to determine if they yield the same results. Several ways of establishing reliability through replication are suggested: inter-subject direct replication, intra-subject direct replication, and systematic replication. Generality can be

established by finding similar results under different conditions. Sidman advocates the use of systematic replication where the experiment is repeated under different conditions instead of direct replication which requires all subjects to be treated alike except for the independent variable in question (Sidman, 1960, p. 111). If similar results using systematic replication are obtained, evidence supporting both reliability and generality is obtained. If systematic replication fails, then the original experiment must be directly replicated; therefore, systematic replication is only sensible when one has enough confidence in the techniques to warrant using the data as a basis for performing new experiments.

Both Campbell and Stanley and Sidman are suggesting similar strategies for using non-experimental designs. Campbell and Stanley suggest the use of the rejection of alternate hypotheses to establish causal relationships, where the designer must be cautious in controlling sources of invalidity. Eliminating these threats to validity increases the strength of the design by eliminating rival hypotheses. Sidman suggests using evidence of generality and reliability resulting from systematic replication to establish causal relationships. The more similar the results found in different settings, the more confidence is gained in establishing causal relationships between variables. Campbell and Stanley (1963, p. 3) also suggest a

need for systematic replication: "The experiments we do today, if successful, will need replication and cross-validation at other times under other conditions before they can become an established part of science, before they can be theoretically interpreted as a part of science."

The designs proposed by these authorities can be thought as having much to offer the formative evaluator. Design specialists caution "because full experimental control is lacking, it becomes imperative that the researcher be thoroughly aware of which specific variables his particular design fails to control" (Campbell and Stanley, 1963, p. 34). Instructional materials specialists (Lumsdaine, 1965; Markle, 1964) stress that any conditions which can affect a program be specified, and either controlled or eliminated. Therefore, formative evaluation can be successful if all factors which can affect the program are specified and causal relationships are established through the elimination of rival hypotheses.

A design or set of procedures which could be useful in formative evaluation would be one concerned with establishing and eliminating rival hypotheses. Platt (1964) also discusses the inappropriateness of classical design methodology of comparing two methods to decide which one yields the more desirable results for many areas of research. Platt describes the use of an "accumulative method of inductive inference" in certain areas of

science, e.g., molecular biology, which has resulted in impressive advances within the field. He has coined the term "strong inference" to describe this method because, in his opinion, the method has been extremely effective in producing rapid progress.

Strong inference is based on the systematic application of inductive inference as introduced by Frances Bacon. Bacon developed a "logical tree" or "conditional inductive tree" which involves listing the alternative explanations of a result. The second step involved designing crucial experiments which exclude one or more of the hypotheses. In the third step, then, the scientist actually performs these experiments. The final step involves recycling the procedure, adding other hypotheses generated from the experiment and systematically eliminating other possibilities. Acceptance of an explanation is based on eliminating all but one alternative. Platt emphasizes that these experiments offer no proofs, only disproofs of other explanations; at any time another explanation could be found which is as good or better than the one previously accepted.

Platt's opinion is that scientists should be designing experiments which disprove rival hypotheses and results should be based on the elimination of alternative explanations. He feels that the use of
in all fields of science allows the scientist to explore the unknown at the fastest rate because there is a minimum

sequence of steps to be followed and conclusions are reached rapidly by eliminating all possibilities except one.

The application of this method of strong or inductive inference to developmental work in instructional system design appears to be a useful alternative to classical design. The use of strong inference would allow the curriculum developer to establish causal relationships by eliminating rival hypotheses rather than seeking the direct cause and effect relationships required by classical designs. The evaluator could be concerned with establishing why instructional materials are not adequate rather than discovering if one set of materials is better than another. An obvious result of establishing why materials are inadequate would be the improvement of the instructional materials as they are being evaluated.

Another advantage in using strong inference in the developmental process stems from the many variables and their interactive effects which affect instructional work in a classroom. The construction of a "logical tree" would require the developer to consider all probable reasons for a result rather than just one. His responsibilities would then include designing experiments to singularly test each probable cause. Results establishing causal relationships would be based on the elimination of all but one of the probable causes. Once a causal relationship had been established through strong inference, it is, as Platt (1964) points out, accepted only until a better explanation

is found. Theoretically the results from the crucial experiment have only disproven other probable explanations.

This theses is an attempt to construct a model for the in-class tryout of an instructional system using Platt's (1964) method of strong inference. The construction of a "logical tree" will consist of listing all probable causes of inadequacies of materials, methods for testing each one, and a basis for deciding whether to reject each cause.

CHAPTER II

THE PROBLEM

This study is concerned with developing and applying procedures for one stage in the design of instructional materials. Nitko (1968) has described three stages in the developmental tryout of lessons and classroom management procedures: (1) a pre-tryout to discover gross inadequacies in the lessons, (2) an in-context developmental tryout to use materials with a representative sample of pupils and classroom conditions, and (3) field testing to use the final revised materials in the intended situation. This study is concerned with the clarification of procedures for the second of these stages, the in-context developmental tryout of instructional systems. During this stage the developmental work of the instructional system is evaluated using a representative sample of pupils and classroom conditions. The curriculum developer and formative evaluator now assess how well the materials work in context, suggest changes, tryout the revisions, and reassess materials. As Nitko (1968, p. 8) states, "The (in-context) tryout refers to a never ending series of experiments." The process of evaluating materials becomes

a continuing one of evaluating materials, identifying and revising poorer materials, and re-evaluating them. The process is continued until an acceptable combination is found.

Most work in curriculum development has not resulted in an acceptable methodology. Although Chitayat (1970) tried three different formative evaluation procedures, none were totally acceptable. Her conclusions did establish a need for evaluating classroom management variables as well as the materials, and emphasized the point that the control of dimensions which effect academic performance was more valuable than collecting masses of difficult to interpret data.

During the formative evaluation of a developing program, the curriculum designer and evaluator seek ways to improve the instructional system. Their task of seeking direct cause-and-effect relationships between instruction and student success can be a slow and often unproductive activity. One of the needs in formatively evaluating an instructional system is a method for making rapid decisions and improvements in the instruction.

It appears useful to apply Platt's (1964) method of strong inference to the area of formative evaluation of instructional systems. Strong inference is based on the exclusion of alternate hypotheses or explanations. Alternatives which cannot be excluded are considered to establish

causal relationships only until they are disproven. At any time another explanation can be found which cannot be disproven. In the area of curriculum development a major problem is acquiring sufficient evidence to prove why materials do or do not work. The number of variables and their interactive effects present in a classroom makes it difficult to tightly control the situation. The application of strong inference as a method for formative evaluation would help in overcoming certain problems created by working in an on-going classroom and providing procedures for making rapid improvements. The evaluator could concentrate on the inadequacies of the materials by asking "why did these materials not work" and listing as alternate hypotheses all variables which potentially could contribute to failure. "Experiments" could then be designed and carried out individually to exclude each alternative. If all listed alternatives are rejected, the evaluator's task then becomes one of seeking other alternative explanations. If one alternative explanation cannot be eliminated, it is momentarily accepted as the "cause" of failure. The instructional system is then modified according to the accepted hypotheses until another failure results, starting the cycle over again.

The problem to be dealt with is the application of "strong inference" (Platt, 1964) to develop formative evaluation procedures which can be used in an on-going

classroom. The procedures will allow for the collection of objective and subjective information which can be used to immediately modify an instructional system and then evaluate these changes. If the formative evaluation procedures are effective in developing instructional materials, then evaluation during the final stage, field testing, should be less extensive.

CHAPTER III

PLAN OF THE STUDY

This study was conducted in two fourth grade classes using the Individually Prescribed Instruction (IPI) mathematics curriculum at Oakleaf School. The curriculum is separated into several levels (A to G), each roughly equivalent to a grade level. At most levels the students work through thirteen different areas (numeration, place value, addition, subtraction, multiplication, division, combinations of processes, fractions, money, time, systems of measurement, geometry, and special topics). Before entering a unit each pupil takes a pretest to determine which objectives within the unit he has or has not mastered. For each objective in the unit, there are self-instructional materials designed to teach the objective and a curriculum embedded test (CET) designed to test mastery of the skill. After all skills within a unit are mastered by scores from either the pretest or CET, a posttest which tests knowledge of all objectives within a unit must be mastered.

This study was conducted over the entire school year in the two fourth grade math classes. Both teachers used in the study had previous experience in using behavior modification techniques to motivate their students and in

writing instructional materials. The curriculum materials evaluated during the year were those the classes would normally be working in.

A. Instructional System Under Study

The curriculum evaluated and modified in this study consists of the objectives in the present IPI mathematics curriculum (levels D, E, F) used by the fourth grade at Oakleaf School. Modifications in classroom management procedures were made to fit the requirements for evaluating the materials. Although the lessons, order of the objectives, and content of the CET's were subject to change in this evaluation process, the content of the objectives, as defined by the posttest, were not altered.

B. Design

In developing this model for in-context formative evaluation major guidance was obtained from examining the procedures described in Platt's (1964) discussion of strong inference. Before using "strong inference" for doing formative evaluation of instructional systems, the evaluator must consider how to adapt the procedures of inductive reasoning to make them appropriate to the needs of curriculum development. Platt (1964) lists four steps which he states must be applied "formally and explicitly and regularly to every problem." The first step involves

considering the alternative explanations for a given result. Platt suggests the use of a structure similar to Bacon's conditional inference tree. In curriculum development it appears necessary to specify what evidence could be used to identify a problem ("a given result"). In the IPI program major evidence of this type would be a test failure. The evaluator then becomes concerned with seeking alternative explanations for the causes of poor test performance. Alternative hypotheses are generated by asking "Why did this occur?" or "What could account for this failure?" Platt's article then suggests devising crucial experiments to test each alternative.

In curriculum development and evaluation various types of evidence may be used to identify a number of similar problems. The listing of alternative hypotheses, and devising crucial experiments is repeated for each such identified problem. Since the evidence used to locate a problem is similar, so are the alternate hypotheses and crucial designs generated to locate the cause. For example, in most situations, the hypothesized reasons for failing a test will be similar regardless of the objective or the test involved. This would seem to make it worthwhile to develop a carefully devised logical structure to study each "type" of evidence used to identify a problem. This logical structure would offer two advantages. First, it would be more generalizable within the curriculum, and thus could be used with all objectives and serve

to eliminate the need for repeatedly devising alternate hypotheses and crucial experiments. Second, this need to apply the same procedure in many specific cases in an on-going classroom makes it important to minimize the number of "crucial experiments" that are performed to locate an acceptable cause of failure. Therefore, Bacon's "logical tree" will be modified by using two steps in generating a logical tree structure. The first involves examining the identified inadequacy (e.g., poor test performance) in such a way as to define the failure as specifically as possible (e.g., failure on "these types of items"). This can serve to delimit the number of explanations of failure that must be tested. The second step involves incorporating into the tree structure certain questions which can be answered by subjectively analyzing materials, pupil performance, etc. The answers to such questions then provide specific suggestions as to which hypotheses should be tested first. In essence, the answers to these questions represent assumptions underlying the related hypotheses.

According to Platt, when a plausible hypothesis is located, a crucial experiment should be designed to exclude the hypothesis. In curriculum evaluation, this would consist of identifying the hypothesis which appears to be the most likely cause of failure and changing some one component of the program to overcome the hypothesized deficiency. Thus each hypothesis would have a specific design associated with it.

The third step then consists of carrying out the experiment to yield results which allow the evaluator to reject or to fail to reject the hypothesized cause of failure. In IPI this would consist of changing one dimension and then retesting the student. If a student passes the test, then the hypothesis cannot be rejected. If he fails the test, the evaluator seeks other possible explanations from the tree and the process is recycled. This recycling procedure is Platt's final step.

There are several advantages in using strong inference in the in-class tryout of materials. One failure by a student given the evaluator input into the evaluation structure. Once the results from a crucial experiment allow the evaluator to fail to reject a hypothesis, an improvement in the materials or classroom procedures can be made immediately. Changes made in the instructional system remain permanent until another student fails a test. No change is permanent, it is only momentarily accepted as the 'cause' of failure until it can be rejected as a possible explanation.

A second advantage of this method should be the rapid improvements that could be made in the program. Failure to reject or rejection of hypotheses can be made quickly and continually within an on-going classroom.

It appears that the end product of such a procedure for the formative evaluation of materials would be

instructional materials with a low error rate. Poor material construction could be identified with a minimum number of students and verified over time with larger numbers.

C. Procedures

This study involved the design and tryout of a specific evaluation technique based on using a quasi-experimental design for establishing cause and effect relationships. The first step involves a careful specification of what was to be evaluated, namely a detailed description of the purposes of the lesson materials and the desired classroom procedures to be followed in the use of the materials. This step is essential to establishing cause and effect relationships. Specifying both of these aspects helps control and identify some sources of invalidity and increases the effectiveness of this technique. This step made use of the four-phase outline, (1) goals, (2) plan, (3) operation, (4) assessment, employed in several other analyses of the IPI development and evaluation process (Lindvall and Cox, 1970).

The second step consists of specifying those evidences of failure in the instructional system which will be taken as indicators that revisions may be necessary. In the examples used in the present study, such evidences of failure are typically poor test performance.

The third step is to generate a list of all possible causes for test failure following the first step of the procedures of strong inference. This will probably require a more detailed analysis of the specific nature of each given failure for many tests. Information regarding what the items missed have in common and what type of error contributed to each item failure will be used to discover all possible causes of test failure. Such information may be helpful in pinpointing weaknesses in lessons, in tests, or in instructional activities. This generating of hypothesized causes of failure will also be aided by an examination of any lesson pages or other types of exercises completed by the student who failed the test. In addition, it will typically be important to observe students as they work on this particular unit of study, to note how they apply themselves and how they self-score their tests. In some cases results from such informal observations will be used to suggest types of controls that may have to be included in the plan to provide for a valid tryout of materials.

The fourth step in the procedure for developing the in-context evaluation process that is the objective of this study is almost a part of the foregoing procedure for generating hypothesized causes of failure. This step requires the integration of such hypotheses into small inference trees selecting all hypotheses which

appear to be probable causes of a specific failure. These trees, in turn, will be integrated into a more comprehensive system for strong inference in the final step of the development.

Following the specification of each of the "small" inference trees described above, experiments must be devised to test each of the proposed hypotheses as required by strong inference. For each hypothesis, a design which changes some one component of the instructional system to overcome each hypothesized deficiency will be generated.

As suggested above, the final step in the procedures used in this study is to develop a comprehensive re-useable inductive inference tree to be used in identifying and relating all of the hypotheses that should be considered in conjunction with each type of evidence of system failure that is of concern. Such a comprehensive tree then provides a basis for evaluating all comparable elements of the instructional system. For example, in the context in which this study was carried out, the IPI math program, these trees are useful not only for evaluating the specific lessons that were the center of attention, but also would be useful for the in-context evaluation of all IPI math lessons.

This comprehensive inference tree, an extension of the tree advocated by Platt, offers two advantages for curriculum development. First, it is more generalizable

within the curriculum because it can be used with all objectives, thus eliminating the need for repeatedly devising alternate hypotheses and crucial experiments. Second, in order to minimize the number of crucial experiments which must be performed, questions which can be answered subjectively by analyzing the student's test and lesson materials will be included in these tree structures. The answers to these questions will provide specific suggestions as to which hypotheses should be tested first, eliminating the need of testing all possible hypotheses.

The procedures outlined in this section provide the structure for the presentation of the results of the study which are given in the following chapter.

D. Data Gathered

In order to use the four-phase outline specified by Lindvall and Cox, information which assesses both the goals and the plan is necessary. Both subjective and objective data must be collected daily in order to assess progress in attaining the specified goals. This includes test results, scores on workpages, information on pupil attention, and information on teacher attention.

The data gathered are used to modify the components of the instructional system in order to insure the goals of the program are met. Subjective data, in the form of classroom observations of the teacher and pupil, are used

to modify the components of the plan. Objective data, resulting from performing crucial experiments, are used to revise the instructional materials.

CHAPTER IV

RESULTS

A. Description of Instructional System

The first step in developing these procedures for the formative evaluation of instructional materials involved a careful specification of the purposes of the lesson materials and the desired classroom procedures. The basic model used for evaluating many other components of the IPI program (Lindvall and Cox, 1969) was used to describe the total instructional system within which the lesson materials are to be used. The IPI model stresses defining the goals, plan, operation, and methods of assessment of a program. The first step in applying the model was to define the goals in terms of how the materials were to function. The components of the plan and operation represent a description of what procedures should be followed by pupils and teachers if the materials are to function as described in the goals. Implementing the plan and operating in the classroom requires continually changing the behaviors of the teacher and student with regard to prescription writing, test taking, work skills, and classroom management until these variables are sufficiently controlled.

The final step, assessment, involves the determination of whether or not the materials are functioning in the way in which the goals state that they should. This assessment should identify specific evidences of failure (e.g., poor test performance) which provide a starting point for the in-context evaluation process. This latter process then proceeds by a systematic analysis of the components of the plan and operation to identify those parts of the lesson materials or other aspects of the program that need to be modified. Specifying and demonstrating the procedure for carrying out this systematic analysis is the purpose of the present study.

Goals

Since these procedures are concerned with developing satisfactory instructional materials, the goals of the evaluation are stated only in terms of the purposes of the instructional materials.

Goal I. The materials should be completely self-instructional; students should be able to learn from the materials without teacher assistance.

Goal II. The materials and all related classroom activities should contribute to mastery of CETs and Posttests the first time they are taken.

Goal III. The materials should be designed and used to encourage self-evaluation skills by the pupils.

Goal IV. The student study skills required by the materials should be identifiable.

Goal V. The materials should be developed for successful use with all students.

Plan

It appears necessary in order to conduct an in-class tryout and evaluation of instructional materials that the evaluator be concerned with the total environment. This includes identifying and controlling the influence of variables which can effect the assessment of the goals.

Previous work concerned with the effects of teacher behavior on student academic performance (Mueller, Light, Reynolds, 1970; Reynolds, Light, Mueller, 1971; Light, Reynolds, Mueller, 1971) had been done by the writer in IPI classes. During this research a list of variables which could effect student performance was generated. This list of variables was obtained from intensive classroom observation of the teacher and the student, from monitoring teacher-pupil interactions, and from studying pupil test results. During this previous research, an attempt was made to study the effects of controlling these variables on pupil performance. Therefore, a list of classroom variables which could effect pupil academic performance in an IPI class was available. The major variables were prescription assignment procedures, test taking procedures, interpretation of test results procedures, classroom management procedures, student behavior skills, and teacher behavior.

In order to decide how to control the potential effects of these variables on pupil performance on the instructional materials, several decisions were necessary. (1) What are practical procedures for dissemination of the program? When developing a total environment, the evaluator should try and design components of the plan which are practical for other classrooms where the materials will be used. (2) The other decision concerning the control of these variables considered ways to evaluate the materials which would not interfere with the normal operation of the classroom.

If the evaluator decided that student performance should not vary as a result of certain variables, two methods were found effective to control the effects of these variables. They can either be eliminated or stabilized. In order to eliminate the effects from a variable, rules were constructed which prohibited the effects. For example, in order to insure that student's test performance was only the result of what was learned from the instructional materials rather than student or teacher assistance, elaborate testing rules were designed. If any rule was broken, the student's test was voided, and an equivalent form was taken.

The other effective method for controlling the effects of some variables was to stablize their effects. Rules and procedures were constructed so that the potential influence of certain variables was systematic and consistent.

For example, teacher behavior is known to influence student performance. The teacher's role in the class was therefore explicitly defined as to what she could and could not do. Although the effects of teacher behavior were not measurable, its potential effects on student performance were consistent.

The effects of certain variables on instructional materials were desirable and, therefore, they were only monitored. Specifically, the effects of student study skills on the instructional materials were not controlled and were always considered a possible cause of error. There were two reasons for only monitoring these skills. It would have been extremely difficult and impractical in an operating class to control each student's study skills. And, the only feasible methods would have interfered with Goal III, developing self-evaluation skills. Therefore, the students were instructed in how to use the materials and the evaluator included methods for assessing the adequacy of the student's skills in using the materials.

The listing of the rules for controlling these variables constitute the plan which can be found in Table 1.

TABLE 1

PLAN FOR THE IN-CONTEXT TRYOUT OF INSTRUCTIONAL MATERIALS

Prescription Assignment Procedures that:

1. require all students to use the same instructional materials.
2. allow students to select the appropriate prescription.
3. allow students who fail a test to receive a new prescription written by the teacher based on the appropriate cause of failure.

Test Taking Procedures which:

1. insure an accurate measure of student performance.
2. forbid any assistance from the teacher, aide, or other students.
3. prevent the student from using the instructional materials during the test.
4. require equivalent forms of tests taken after each test failure.

Test Interpretation Procedures which:

1. provide an accurate decision about mastery of each objective.
2. are consistent across students and tests.
3. define tests as the standard of performance.

Classroom Management Procedures which:

1. encourage students to learn from the materials.
2. provide for student decisions.
3. decrease the amount of down time.
4. are consistent.

TABLE 1--Continued

PLAN FOR THE IN-CONTEXT TRYOUT OF INSTRUCTIONAL MATERIALS

Student Behaviors which:

1. permit self-scoring of materials.
2. allow students to be self-evaluators.
3. allow students to solve their own problems.

Teacher Behaviors which:

1. use reinforcement techniques to motivate students.
2. prohibit student tutoring.
3. provide consistent behavior day to day.
4. provide consistent judgements of student performance.

Operation

The third component of the Lindvall and Cox model, the operation, should describe how each part of the plan should look during the evaluation of the materials. As was previously stated, each of the components in the plan were either stabilized, eliminated, or monitored through the use of rules. These rules can be found in Appendix A. Both informal and formal classroom observations were taken daily in the operating classroom during the entire evaluation of the instructional materials. The formal observations were concerned with the teacher and the number of interactions with each pupil during the class. An attempt to maintain a relatively consistent number of interactions was considered important.

Informal observations were made by the evaluator for two purposes. The main reason was to insure that all designated rules were consistently followed by the teacher and student. If the observer noted any breakdown in the rules, nothing was done until after class when the teacher was requested to reinforce students for following the rules. During the evaluation of the instructional materials used by a student, this informal information was considered. For example, if a student was observed misbehaving in the testing area and then failed his test, the evaluator might consider "non-attending to task" while taking a test as a possible reason for failure.

The other result of informal observations of the operating classroom was the chance to observe other variables which could effect academic performance or other rules which could improve control of variables within the classroom. For example, students were required to exchange their pencils for a red pen before they scored their CETs in order to prevent students from changing answers. During the year, the observer noticed a student using the eraser of the red pen to erase an inappropriate digit in an answer. The erasers were then removed from the red pens.

The result of using formal and informal observations was to insure that the conditions specified by the plan were in actual operation. All discrepancies were noted and considered by the evaluator during the evaluation

of the materials. More importantly, the observations were used to maintain the rules dictated by the plan.

Methods of Assessment

The final component in the Lindvall and Cox model, assessment, describes methods for determining how well each goal is achieved. In the outline given below the assessment column described those conditions that should exist if the instructional materials are functioning properly. Since formative evaluation as defined in this paper includes making improvements, the evaluator's task is to identify any needed modifications in the instructional environment so that the goals can be met. This includes changing the way in which lessons are used as well as changing the lessons themselves.

Goal	Assessment
I. The materials should be completely self-instructional.	The teacher is not allowed to tutor students. The success or failure of the materials is based on how well they are self-instructional.
II. The materials and all related classroom activities should contribute to mastery the first time.	a. Mastering of tests the first time should be consistently reinforced by the teacher. b. The materials are evaluated against the standard of passing the first time only.
III. The materials should be designed and used to encourage self-evaluation skills by the student.	a. Materials should be designed so the last page is equivalent in format and content to tests. Students are encouraged to not take a test if they could not do this page correctly.

Goal	Assessment
	b. Students are given the freedom to decide when they were ready for a test. If a cause of failure is identified as poor self-evaluation skills, students are reinforced in good skills. c. Students are reinforced by the teacher and the system for good self-instructional skills.
IV. Student study skills required by the materials should be identifiable.	a. When cause of failure is identified as poor study skills, students are instructed in how to use proper skills. b. If the required study skill is beyond the student's capabilities, the materials are revised.
V. Materials should be developed for successful use with all students.	Materials are continually revised to be successful with everyone. The object is to find one path between the objective and the test. (Once a path is established, other parts or shorter paths can be established.)

The objective of the methods of assessment is to meet the goals. By defining the plan, certain variables which could interfere with the goals can be controlled. Two main sources for not meeting the goals are the instructional materials, which include the objective, the materials, and the tests, and the student's use of the materials. The purpose of the in-context tryout is to develop, revise, and evaluate the materials until they meet these goals.

B. The Identification of Inadequate Materials

The second step involved in developing these procedures was the systematic identification of inadequate materials. For the materials evaluated in this study, objectives that were a part of previously validated hierarchies were used. The materials also contained two bases for determining mastery of each objective, a CET taken after completing each objective, and a unit posttest taken after completing several objectives of similar content. If pupil test performance was not equivalent on the two tests, the unit posttest was always accepted as the mastery criterion. Since students completed these tests frequently, these formative evaluation procedures were designed to use poor pupil test performance to identify inadequate materials.

The major assumption in the design of these procedures was that the "cause" of any poor test performance could be identified by systematically examining pupil performance on instructional materials. Once a possible cause of failure is located, new materials are designed to eliminate the hypothesized cause. Student performance on an equivalent test is then used to assess the revised materials. The evaluator always tests his revisions objectively. The use of an equivalent form of the failed test to assess the revisions provides the evaluator with immediate feedback on his success in hypothesizing a cause

of failure. If an inappropriate cause of failure is used as a basis for revisions, the student should not pass the equivalent test, forcing a repetition of the entire process.

After each class all tests indicating less than perfect pupil performance were analyzed to identify possible inadequate materials. All materials used by the student to learn the objective were gathered for systematic examination by the evaluator and teacher. For each test failure, the question was asked, "Why did this student fail the test associated with these materials?" The successful use of these procedures if based on systematically locating and testing each hypothesized cause of failure.

C. Possible Causes of Test Failure

The next step involved in developing these procedures was to generate a test of probable causes of test failure. In order to discover these probable hypotheses, these five questions were always answered by the evaluator in analyzing the cause of failure.

1. What was similar about the items missed on the test?
2. How did the items missed differ from those items passed on the test?
3. Where in the instructional materials were these types of items presented?

4. What in the instructional materials caused the students to fail the test?
5. How can the hypothesized cause of failure be experimentally proven?

These five questions were answered for each test failure. Once a probable hypothesis was located, a crucial experiment was designed and carried out to test the hypothesis.

The result of this procedure was an extensive list of probable causes of test failure on CETs and Posttests. An example of such a list of causes is represented by the following hypotheses.

If a pupil has failed an objective on the posttest

a. and passed the objective on the pretest, then the pretest and posttest may not be parallel forms.

b. and passed the CET, then the CET and posttest may not be equivalent in either form or content.

c. and passed the CET, then the prescription may not provide enough practice for learning to occur.

d. and passed the CET, then the pages and CET may not teach him how to discriminate directions.

e. and passed the CET, then he may not have sufficiently reviewed before taking the test.

f. and passed the CET, then he may not have checked over his work.

g. and passed the CET, then the criterion for mastery performance may not be adequate.

h. then he may not be motivated to pass the test.

i. then he may not have been "attending to task" while taking the test.

If a pupil fails a CET, then

a. the pages may not teach and provide practice on the tested content.

b. the pages may not teach and provide practice on "unique" properties.

c. the pages may not require adequate practice.

d. the prescription may not contain pages which are duplicates in form and content of the CET.

e. the prescription may be inadequate.

f. the pages may not provide practice involving the same format as the test.

g. he may not have learned from the teaching pages.

h. his work may demonstrate poor work skills.

i. he may have done the prescription incorrectly.

j. he may not have the appropriate prerequisite behaviors.

k. he may not be motivated to do accurate work.

l. he may not be "attending to task" while doing his work.

m. he may not be checking his work.

n. he may not be able to use self-evaluation skills to decide if he has learned the required skills.

D. Construction of Small Inference Trees

As stated previously the analysis of each test failure resulted in a small inference tree which listed several probable hypotheses to be tested. The evaluator then chose one hypothesis and designed and carried out a crucial experiment. The result of each such experiment was that the student either passed or failed an equivalent test. If the student passed this test, the hypothesis was tentatively confirmed and revisions were accordingly made in the materials or plan of the program. If the student failed this test, the hypothesis was rejected and another hypothesis as to the cause of failure was tested.

The use of these steps to locate problems and to improve materials or procedures can be explained most clearly through the use of examples. For each example a copy of the student's test is always presented. In all examples, the handwritten responses are the student's answers. Those marked with an X are incorrect. These tests were taken after the student had used the appropriate mathematics lesson material.

Example I

Figure I shows a pupil's responses to a four-item test on multiplication, where the student missed three questions.

Write in the missing numbers using the associative principle.

$(4x2)x5 = 4x(2x\underline{5})$

$= 4x\underline{10}$

$= \underline{40}$

$2x(4x8) = (2x4)x\underline{8}$

X $= \underline{32}x\underline{8}$

$= \underline{256}$

$(9x3)x6 = 9x(3x\_\_)$

X $= \underline{27}x\underline{18}$

$= \_\_\_$

$6x(7x4) = (6x7)x\underline{4}$

X $= \underline{28}x\underline{42}$

$= \_\_\_$

FIGURE I

A STUDENT'S RESPONSES TO A FOUR-ITEM TEST ON MULTIPLICATION

After class the student's test and materials were gathered for analysis, and answers were sought for the first two of the questions stated previously.

1. What was similar about the items missed on the test?
   a. The student always made the first error on the second line of the problem.
   b. The errors appear to be systematic. The pupil always puts the product of the multiplication problems within both sets of parentheses from the first line into the blanks on the second line.

2. How did the items missed differ from those items passed on the test?

   a. The one item passed had one numeral, a 4, already written in the second line.

Because the single problem passed by the student contained an additional cue, the numeral four, the evaluator hypothesized that the student probably had not learned what the associative principle was from the instructional materials, even though he passed the one item.

The model requires the evaluator to then look through the student's lesson materials to identify a probable reason why the student did not learn the appropriate skill. These instructional materials were desinged so that the last page before a test is equivalent in content and format to the test. Since the pages, upon inspection, appeared to explain the associative principle clearly and the student had completed the pages correctly, attention was focused on the last page before the test. Examination of the last page, appearing in Figure II, led to a hypothesis on the cause of failure, based on the third of the five questions.

3. Where in the instructional materials were these types of items presented? What are possible inadequacies in this presentation?

   a. The format on this page differed from the test. The student was required to use the product of the numerals within the

parentheses on the first line to fill in the blanks on the second line.

Multiplication is associative:

$$(8x2)x2 = 8x(2x2)$$

$$16 \times 2 = 8 \times 4$$

$$32 = 32$$

Write in the missing numbers and solve the equation using the associative principle:

$$(3x2)x5 = 3x(2x\_\_\_)$$

$$\_\_\_\_x5 = 3x\_\_\_\_$$

$$\_\_\_\_ = \_\_\_\_$$

$$(3x9)x4 = 3x(\_\_\_x4)$$

$$\_\_\_\_x4 = 3x\_\_\_\_$$

$$\_\_\_\_ = \_\_\_\_$$

$$(7x6)x3 = 7x(6x3)$$

$$\_\_\_\_x3 = 7x\_\_\_\_$$

$$\_\_\_\_ = \_\_\_\_$$

FIGURE II

THE LAST PAGE WORKED BEFORE THE TEST IN FIGURE I

b. The student also always had an arrow to aid him in putting the product in the correct place.

c. This page also differed from the test in that the student solved each problem for both forms of the equation (axb)xc and

ax(bxc). On the test he was required to solve only one side of each equation, thus eliminating a check of his work.

Once the evaluator has identified differences between the materials and the test, he must choose one possible cause of failure. If an inappropriate cause is hypothesized, student performance should not improve and the evaluator will have to select another cause. This involves answering questions 4 and 5.

4. What in the instructional materials caused the student to fail the test?

   Hypothesis to be tested:

   If the last page of the materials is changed to include problems similar in format to those on the test, then the student will pass the test.

5. How can the hypothesized cause of failure be experimentally proven?

   The following page, shown in Figure III, was added as the last page in the materials. The page does not use arrows to indicate where the products should be placed and responses to one side of the equation are required.

After the student completed this page, he passed an equivalent test. The revised last page was then

included in the materials for all students. No further evaluation of these materials will occur until another student fails the same test.

Solve each equation:

(2x5)x3 = 2x(5x3)
= ___x___
= ____

(3x1)x2 = 3x(1x2)
= ___x___
= ____

(2x7)x3 = 2x(7x3)
= ___x___
= ____

(8x1)x3 = 8x(1x3)
= ___x___
= ____

(3x5)x6 = 3x(5x6)
= ___x___
= ____

FIGURE III

THE REVISED PAGE ADDED TO THE MATERIALS

There is no way for an evaluator to "know" conclusively if his hypothesized cause of failure is correct. If student have no further trouble with the materials

and tests, the formative evaluation has improved the materials. If the same student or another student continues to have trouble, the formative evaluation has not located the problem.

Example II

In the first example, the addition of a new last page was sufficient for that student to achieve mastery of the objective. It was not sufficient for another student, as illustrated by the test in Figure IV.

Write in the missing numbers using the associative principle.

A

(4x2)x5 = 4x(2x 5)

= 4x 10

= 40

B

2x(4x8) = (2x4)x 8

X = 8 x 32

= 256

C

(9x3)x6 = 9x(3x 6)

= 9 x 18

= 162

D

6x(7x4) = (6x7)x ___

= 42 x 28

X = 70

FIGURE IV

ANOTHER STUDENT FAILS THE SAME TEST

After this student failed the test, the same procedures used in the first example were repeated. The student's test and materials were gathered for examination,

and the same questions were answered.

1. What was similar about the items missed on the test?
   a. Both problems missed were of the form (axb)xc.
   b. The student's errors on the second line were systematic. The incorrect answers on the second line were a result of multiplying (axb) and (bxc).
   c. The student's errors on the third line were different. In problem B he multiplied the numerals in line 2, in problem D he added the numerals in line 2. The difference in the error on line 3 was considered of lesser importance by the evaluator because the student had not previously learned how to multiply two two-digit number, which could account for his adding instead of multiplying the numerals.
   d. Both items missed were on the right column on the page.
2. How did the items missed differ from those items passed on the test?
   a. The items passed were of the form ax(bxc), the items failed were of the form (axb)xc.

b. Both items passed were on the left column of the paper.

3. Where in the instructional materials were these types of items presented?

For reasons similar to those discussed in the first example, the evaluator focused his attention on the new last page, presented in Figure III.

a. The page was done correctly by the student.

b. All the problems on the page were of the form ax(bxc).

4. What in the instructional materials caused the student to fail the test?

Hypothesized cause of failure:

If the last page of the materials is revised to include practice in doing both forms (axb)xc and ax(bxc) of the associative principle, then the student will pass the test.

5. How can the hypothesized cause of failure be experimentally proven?

The last page of the materials was again revised to include problems of both forms ax(bxc) and (axb)sc of the associative principle.

Solve each equation:

(2x5)x3 = 2x(5x3)

= ___x___

= _____

3x(1x2) = (3x1)x2

= ___x___

= _____

(2x7)x3 = 2x(7x3)

= ___x___

= _____

8x(1x3) = (8x1)x3

= ___x___

= _____

3x(5x6) = (3x5)x6

= ___x___

= _____

FIGURE V

THE REVISIONS OF THE NEW REVISED PAGE

Once this revised page, shown in Figure V, was used by this student, he was given an equivalent form of the test. He and his fellow students had no further trouble with this objective during the year.

Example III

The first two examples were chosen to illustrate not only how tc use these procedures to identify inadequate materials but to demonstrate how student test performance can be used to continually evaluate the evaluator's decisions. No original instructional materials or revisions

are ever free from the possibility of undergoing revision.

	Skip count by 3's.							
A	472,	475,	478,	471,	474,	477,	490	X
B	205,	202,	299,	296,	293,	290,	187	X
C	747,	750,	753,	756,	759,	762,	765	
D	1,000,	997,	994,	991,	998,	995,	982	X
E	638,	641,	644,	647,	650,	653,	656	

FIGURE VI

A STUDENT'S RESPONSES TO A FIVE-ITEM TEST ON NUMERATION

The third example has been selected to illustrate how these procedures can identify a wide range of causes of failure. This illustration demonstrates why an evaluator must consider not only current instructional materials but also the prerequisite materials. In Figure VI is a student's test for an objective requiring the student to skip count by 3's. Again the evaluator's questions are raised.

1. What was similar about the items missed on the test?
    a. The pupil's errors in skip-counting are always made when the pupil has to change the place value in the tens or hundreds place.
2. How did the items missed differ from those items passed on the test?

a. The pupil can skip-count by 3's when the place values for some multiples of 10 but not for all.

3. Where in the instructional materials were these types of items presented? How did the student respond to these items?

In looking over the student's materials, the pupil consistently made errors when he had to change place values and there were no clues about the value of the new place value. In line C, the last numeral, 765, could provide a clue as to what the new place value should be. Since the materials were designed to teach skip-counting by 3's and the student did demonstrate he could skip-count without changing place values, the evaluator decided to consider the prerequisite behaviors. The immediate prerequisite behavior required that the student be able to count by 1's. The criterion test for this behavior is presented in Figure VII. Since the student's failed test indicated a possible problem in counting by 1's, the test for this behavior was examined. The series in line A was the only series which required the student to name the next place value without any clues except for line B which was considered,

because of prior experience, easier for students to learn. Line D and G, although they required a change in place value, also provided additional clues as to what the new value should be. If a student only missed line A, it would be possible to be given mastery if the evaluator, at that time, was not aware of the uniqueness of this line.

	Count forward by 1's:
A	37 , 375, 376, 377, 378, 379, 370, 371 X
B	995, 996, 997, 998, 999, 1000
C	230, 231, 232, 233, 234, 235, 236
D	659, 660, 661, 662, 663, 664, 665, 666
	Count backward by 1's
E	529, 528, 527, 526, 525, 524, 523
F	837, 836, 835, 834, 833, 832, 831
G	311, 310, 309, 308, 307, 306, 305

FIGURE VII

THE STUDENT'S RESPONSES TO THE SEVEN-ITEM TEST OF THE PREREQUISITE BEHAVIORS IN NUMERATION

4. What in the instructional materials caused the student to fail the test?

   Hypothesized cause:

   If a student cannot count by 1's to 1000, then he cannot skip count by 3's correctly.

5. How can the hypothesized cause of failure be experimentally proven?

   A revised test for the prerequisite behavior in numeration was constructed to include more place value changes without providing clues about the new place value. This student was not able to master this test, so he was reassigned the materials to teach him how to count to 1000.

The revised test was substituted into the curriculum. No student who mastered the revised test failed the test of skip counting by 3's.

This example illustrates how the lack of prerequisite skills can cause inadequate performance in future objectives. This illustrates why an evaluator must consider all aspects of an instructional system which can effect student performance not only the failed test and its associated materials.

Example IV

When a pupil fails a test, there may be nothing wrong with the instructional materials. The problem

could be the way in which the materials were used by the student. Our instructional materials were designed to be used in a specified way. Evidence indicated that deviations from the specified procedures could result in inadequate test performance for some pupils. The evaluators always had to consider inappropriate work skills as a potential cause of failure which could be tested. An example of this can be found in analyzing the cause of failure for the test in Figure VIII.

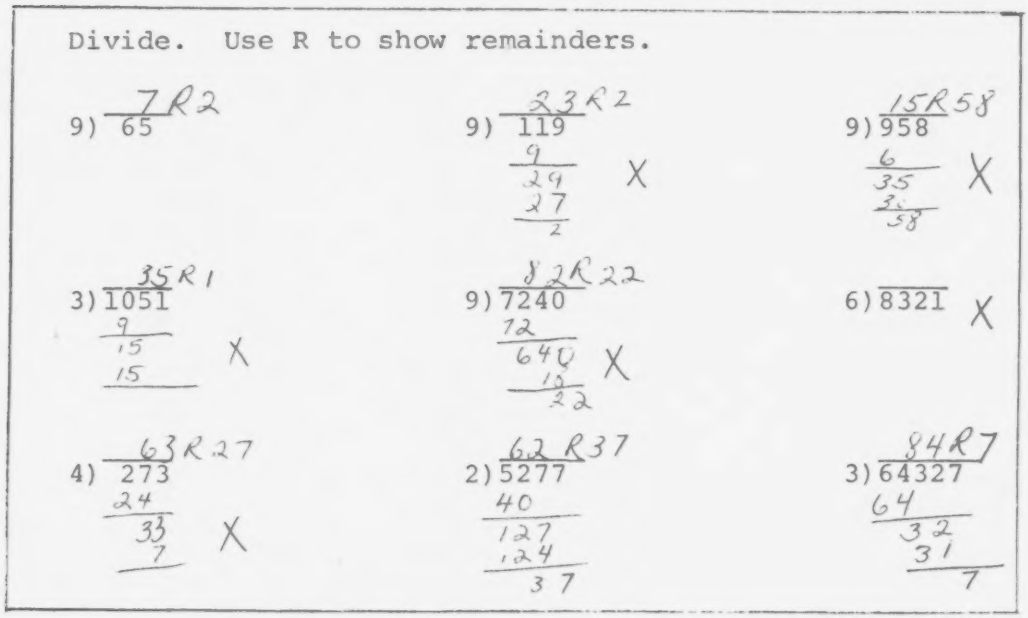


FIGURE VIII

A STUDENT'S RESPONSES TO A NINE-ITEM TEST ON MULTIPLICATION

Note the answers to the evaluator's questions in this case.

1. What was similar about the items missed on the test?

   a. The errors do not appear to be systematic.

1.
```
  23 R2
9)119
  9
  29
  27
   2
```
The student put the remainder of 2 as the first digit in the quotient or he multiplied 9x2=9. He finished the problem correctly.

2.
```
  35 R1
3)1051
  9
  15
  15
   1
```
The student was correct until he got to the third digit in the quotient, where he left out the 0 before stating the remainder.

3.
```
  62 R37
2)5277
  40
  127
  124
   37
```
There does not appear to be any reason for his responses to this problem.

   b. The student's answers are always two-digit numbers with a remainder.

   c. There were many erasures on the test which could indicate confusion.

2. How did the items missed differ from those items passed on the test?

   a. The only problem done correctly had a two-digit dividend.

Find the quotient and remainder for each example.

$6\overline{)1551}$	$4\overline{)2714}$	$2\overline{)1972}$
$3\overline{)2915}$	$7\overline{)6362}$	$6\overline{)5830}$

FIGURE IX

THE LAST PAGE COMPLETED BY THE STUDENT BEFORE TAKING THE TEST IN FIGURE VIII

3. Where in the instructional materials were these types of items presented?

   The student's workpages were examined. The last page of the materials, shown in Figure IX, was done perfectly by the student. The page contains quotients with more than two digits and the pupil never had a remainder greater than the divisor. The items on the page also consisted of more than two-digit dividends.

Because students were allowed to score their own workpages, they had access at all times to the answers to their workpages. Since this student only answered one problem correctly on the test, since his errors were

inconsistent, but his workpage responses were perfect, the evaluator had to consider the possibility that the student misused the answer keys.

4. What in the instructional materials caused the student to fail the test?

   Hypothesis:

   If a student uses an answer key to copy answers on his instructional materials, then he will fail the test.

5. How can the hypothesized cause of failure be experimentally proven?

   The student was reassigned the identical materials except that the teacher scored his workpages. If the student had not used the key incorrectly, he should be able to do the pages correctly again and still fail the test. He will no longer have a key to provide answers for his pages. If the materials are faulty, he will still fail the equivalent test; if the materials are adequate, he should pass the test.

The student had difficulty in re-doing his materials. After he completed the assigned materials, he passed an equivalent test. Since no other student during the year failed the test, the evaluators felt their hypothesis of poor self-evaluation practices was correct.

E. Comprehensive Inference Trees

The preceding examples were chosen to illustrate how the causes of inadequate student test performance can be identified and corrected. In order to establish cause and effect relationships between test failure and instructional materials, the four steps of Platt's strong inference were used. After each student failed a test, a hypothesis proposing a cause and effect relationship was selected. The component of the instructional system identified by the hypothesis as inadequate was changed and the student was re-tested on an equivalent form of the failed test. If the student passed an equivalent test, the hypothesis was accepted as identifying the cause of test failure. If any revision was made in the instructional materials, it was included in the system for use by all students.

The procedures for locating possible causes of each test failure resulted in a "mini-inference tree" consisting of all probable hypotheses which could be tested to establish cause and effect relationships for each test failure. After using these procedures with many tests, three ways to improve the efficiency of this process became apparent: (1) Because the procedures were designed for use in an on-going class, it was impractical to always individually test all probable causes of test failure. Therefore, it became apparent that developing methods for systematically selecting first the most appropriate

hypothesis to be tested would increase the efficiency of these procedures. (2) There are a finite number of probable hypotheses that can be tested to establish cause and effect relationships for all tested objectives. Efficiency of these procedures could be increased by generating generalizable inference trees that could be used with all objectives in the curriculum. (3) Through experience, it was found that the task of identifying the cause of each test failure could be simplified by analyzing the specific type of failure the student made on all test items and lesson pages. Certain types of failures were found to be more likely to be explained by certain hypotheses than by others. A series of questions was designed whose answers could lead the evaluator into selecting the more appropriate hypothesis first. These questions included such things as: how many test items did the student fail? Were the test errors systematic? Were the test errors computational or process?

The number of items a student fails on a test can also offer information to the evaluator as to the cause of test failure. If a student fails one or two items, the type of error is usually the result of either a process error or a computational error. A process error is defined as an error in an item resulting from the student not learning the exact process being taught by the materials. When a student only misses one or two problems

because of a process error, it usually means that the items are unique in their content. For example, a student only failed the items on a test in subtraction with borrowing which contained a zero in the tens place but passed all other subtraction items. The probable cause of failure can be found by analyzing the content of the items failed and the items passed to note how they differ. Once a uniqueness has been identified, the evaluator can use a branch of the "inference tree" to choose a testable hypothesis.

A computational error is defined as an error in an item resulting from the student writing the incorrect sum, product, quotient, or difference in a paoblem. When a student only misuses one or two problems because of a computational error, it usually means the student needs practice in these prerequisite skills or he needs to be reinforced for accuracy.

If a student fails more than two items on a test, the evaluator should decide if the errors are systematic. Systematic errors are defined as errors resulting from the student using the identical but incorrect rule to answer all items. An example of a systematic error can be found in the first example. The student always wrote the products of the numerals within the parentheses from the first line on the second line. Once the evaluator determines the rule the student is using to yield his

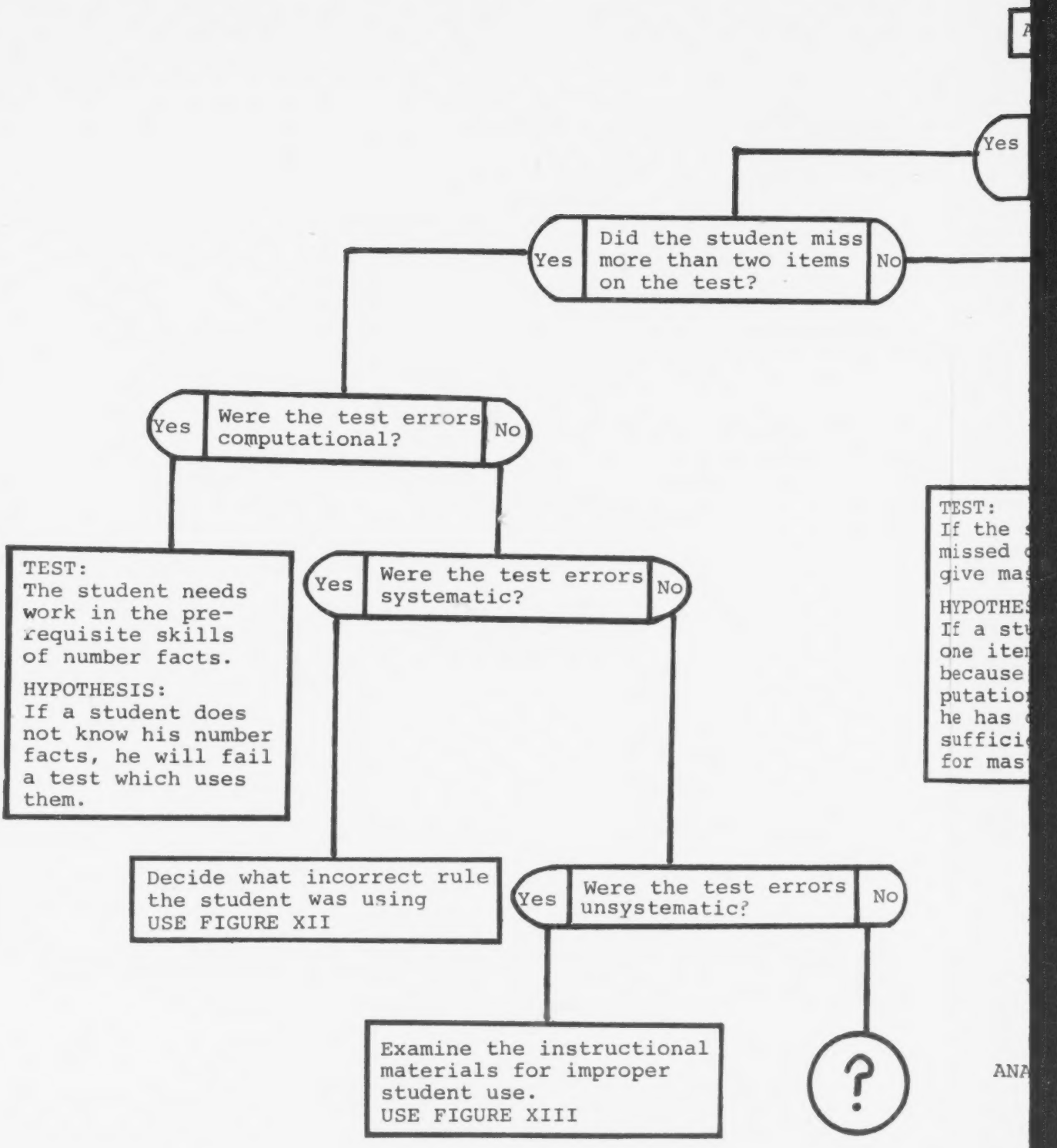
incorrect answers, he can use a branch of the "inference tree" to choose a testable hypothesis.

Errors which are non-systematic are those which usually result from the student answers problems incorrectly for different reasons, as illustrated by the last example. Once the evaluator determines that errors are non-systematic, this leads him to a part of the inference tree where he can choose a testable hypothesis.

The results of answering these questions about failed test items can usually aid the evaluator in immediately eliminating many probable hypotheses of the causes of test failures.

These first parts of the trees, shown in Figure X and XI, includes those questions which can be answered by examining student performance on CET and Posttest items. The answers to these questions permit the evaluator to eliminate certain further branches of the trees and direct his attention to those branches that should be pursued first.

The second section of each tree consists of several branches found in Figures XII, XIII, XIV, XV, XVI, XVII, and XVIII, consisting of specific testable hypotheses related to the different causes of test item failure. Each of these branches also contains a series of questions the evaluator can answer about the student's use of the instructional materials. The answers to these


Yes
Did the student miss more than two items on the test?
No
Yes
Were the test errors computational?
No
TEST:
The student needs work in the pre-requisite skills of number facts.
HYPOTHESIS:
If a student does not know his number facts, he will fail a test which uses them.
Yes
Were the test errors systematic?
No
Decide what incorrect rule the student was using
USE FIGURE XII
Yes
Were the test errors unsystematic?
No
Examine the instructional materials for improper student use.
USE FIGURE XIII
?

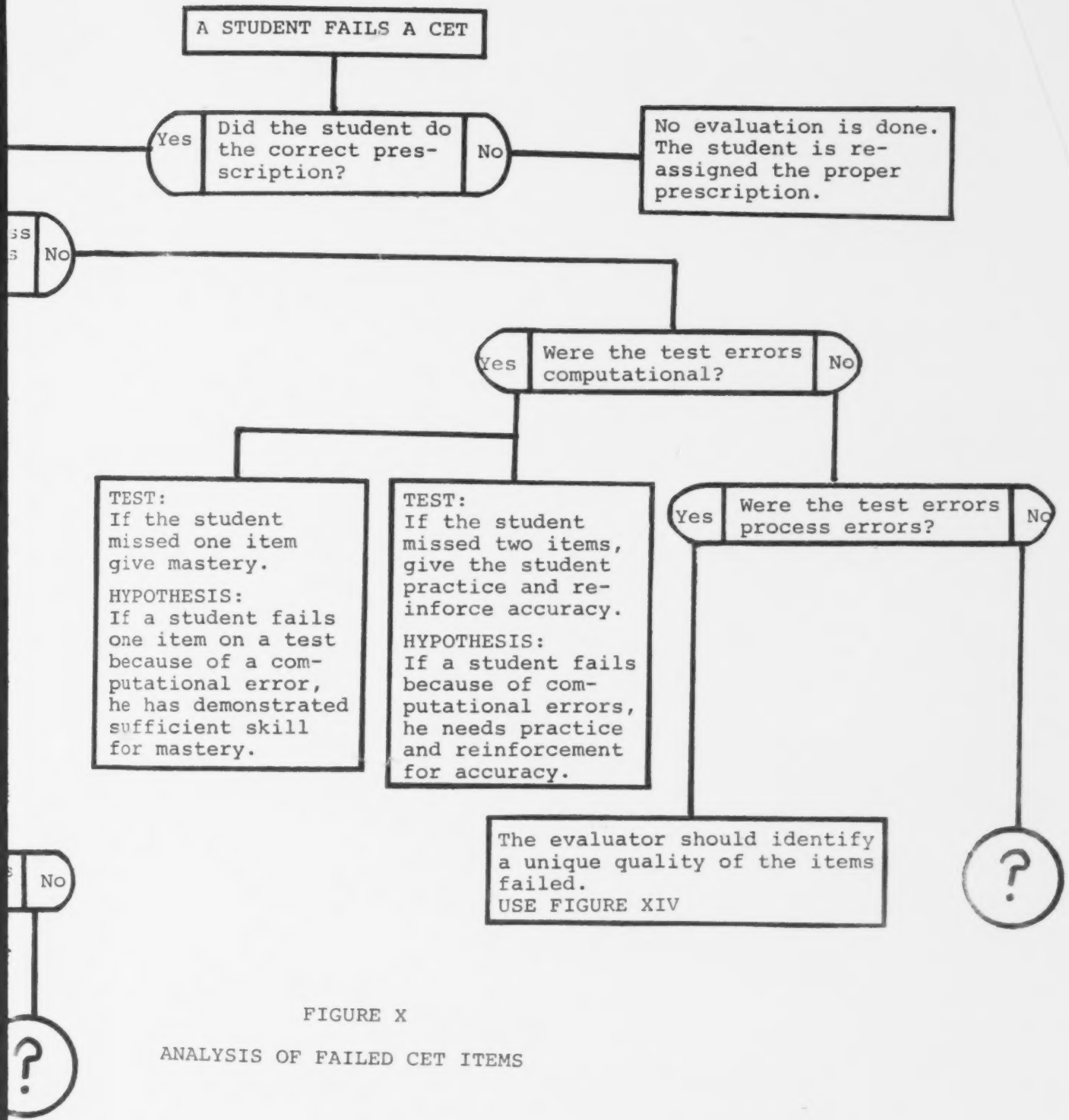

A STUDENT FAILS A CET
Yes
Did the student do the correct prescription?
No
No evaluation is done. The student is reassigned the proper prescription.
No
Yes
Were the test errors computational?
No
TEST:
If the student missed one item give mastery.
HYPOTHESIS:
If a student fails one item on a test because of a computational error, he has demonstrated sufficient skill for mastery.
TEST:
If the student missed two items, give the student practice and reinforce accuracy.
HYPOTHESIS:
If a student fails because of computational errors, he needs practice and reinforcement for accuracy.
Yes
Were the test errors process errors?
No
The evaluator should identify a unique quality of the items failed.
USE FIGURE XIV
?
No
?

FIGURE X

ANALYSIS OF FAILED CET ITEMS

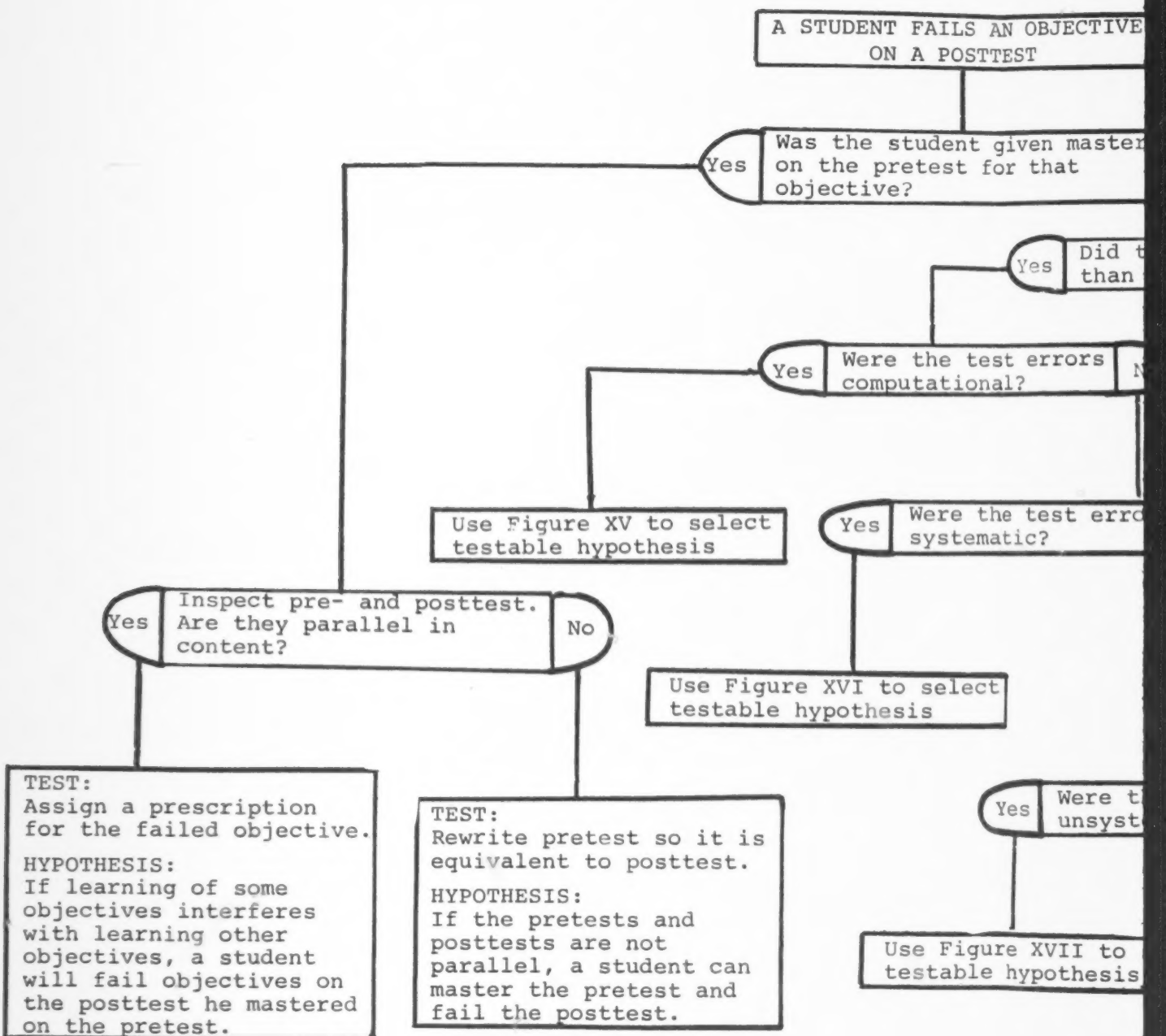

A STUDENT FAILS AN OBJECTIVE ON A POSTTEST
Was the student given master on the pretest for that objective?
Yes
Did t than
Yes
Were the test errors computational?
N
Use Figure XV to select testable hypothesis
Yes
Were the test erro systematic?
Inspect pre- and posttest. Are they parallel in content?
Yes
No
Use Figure XVI to select testable hypothesis
TEST:
Assign a prescription for the failed objective.
HYPOTHESIS:
If learning of some objectives interferes with learning other objectives, a student will fail objectives on the posttest he mastered on the pretest.
TEST:
Rewrite pretest so it is equivalent to posttest.
HYPOTHESIS:
If the pretests and posttests are not parallel, a student can master the pretest and fail the posttest.
Yes
Were t unsyst
Use Figure XVII to testable hypothesis

FIGURE XI

ANALYSIS OF FAILED POSTTEST

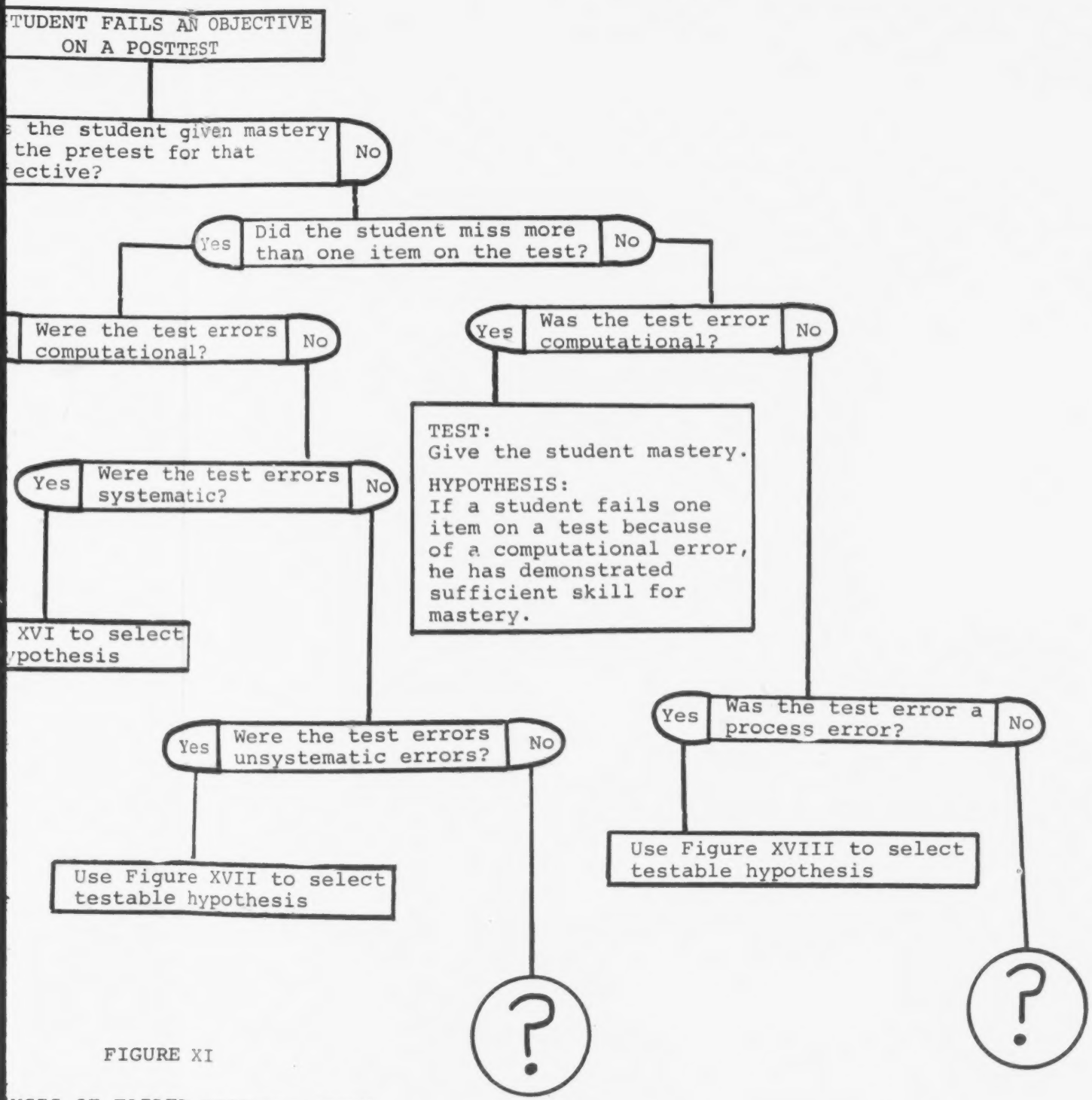


FIGURE XI

YSIS OF FAILED POSTTEST ITEMS

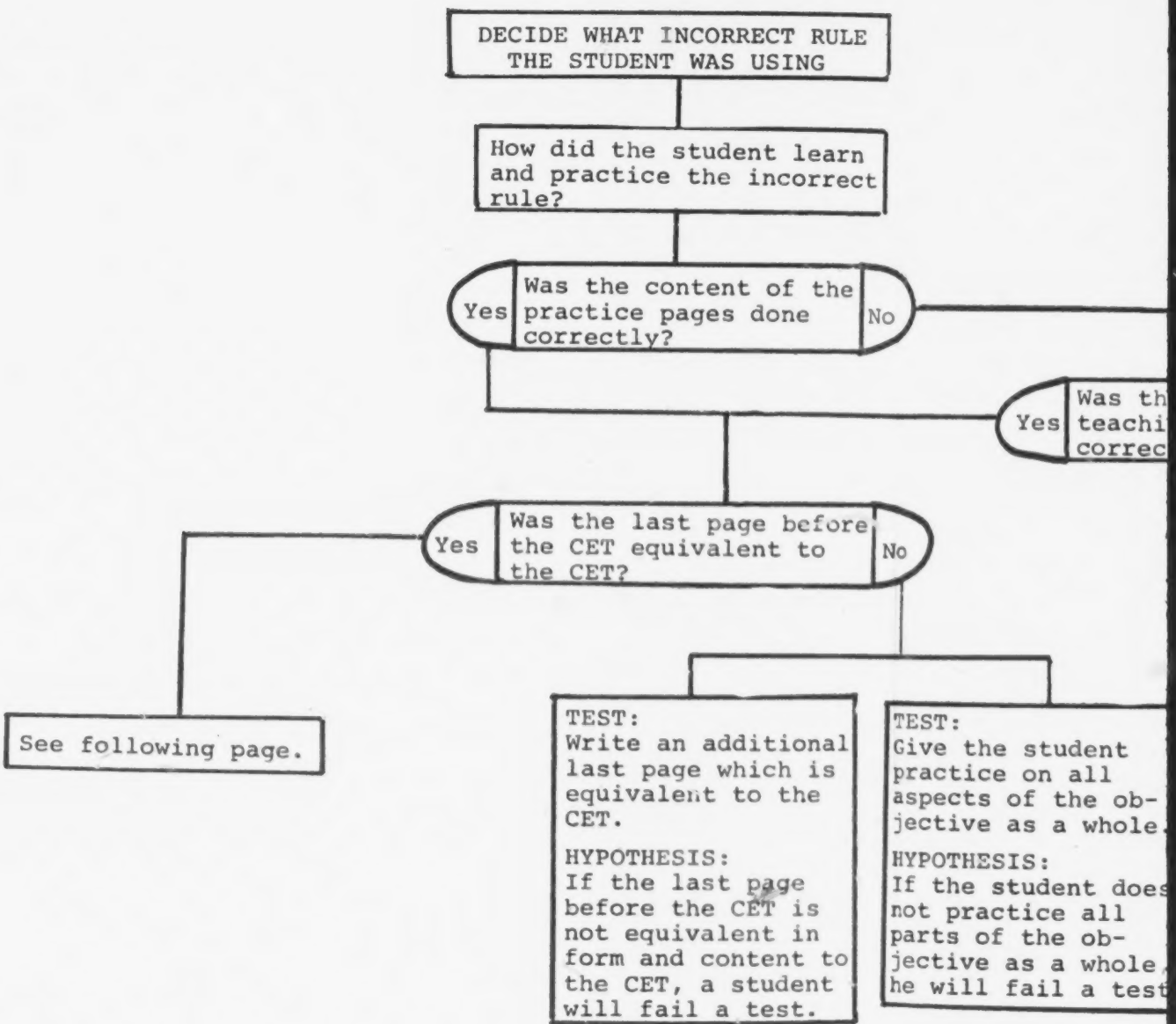

DECIDE WHAT INCORRECT RULE THE STUDENT WAS USING
How did the student learn and practice the incorrect rule?
Yes
Was the content of the practice pages done correctly?
No
Yes
Was th teachi correc
Yes
Was the last page before the CET equivalent to the CET?
No
See following page.
TEST:
Write an additional last page which is equivalent to the CET.
HYPOTHESIS:
If the last page before the CET is not equivalent in form and content to the CET, a student will fail a test.
TEST:
Give the student practice on all aspects of the ob-
jective as a whole.
HYPOTHESIS:
If the student does not practice all parts of the ob-
jective as a whole, he will fail a test

FIGURE XII

ANALYSIS OF INSTRUCTIONAL MATERIALS FOR STUDENT USING INCORRECT RULE

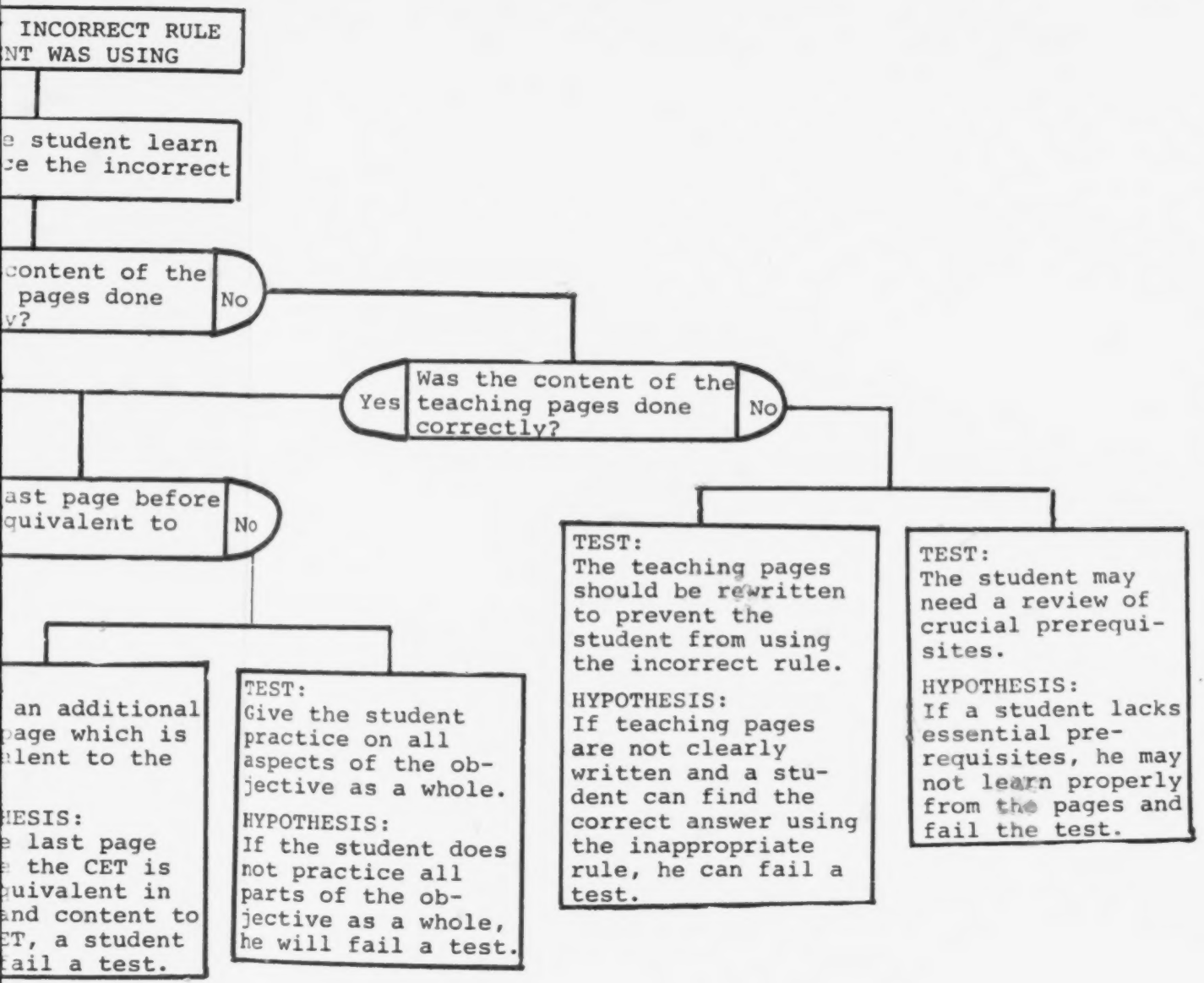

INCORRECT RULE
NT WAS USING
e student learn
ce the incorrect
content of the
pages done
v?
No
Yes
Was the content of the teaching pages done correctly?
No
ast page before
quivalent to
No
an additional
page which is
alent to the
ESIS:
e last page
the CET is
uivalent in
nd content to
T, a student
ail a test.
TEST:
Give the student practice on all aspects of the objective as a whole.
HYPOTHESIS:
If the student does not practice all parts of the objective as a whole, he will fail a test.
TEST:
The teaching pages should be rewritten to prevent the student from using the incorrect rule.
HYPOTHESIS:
If teaching pages are not clearly written and a student can find the correct answer using the inappropriate rule, he can fail a test.
TEST:
The student may need a review of crucial prerequisites.
HYPOTHESIS:
If a student lacks essential prerequisites, he may not learn properly from the pages and fail the test.

FIGURE XII

RUCTIONAL MATERIALS FOR STUDENT
NG INCORRECT RULE

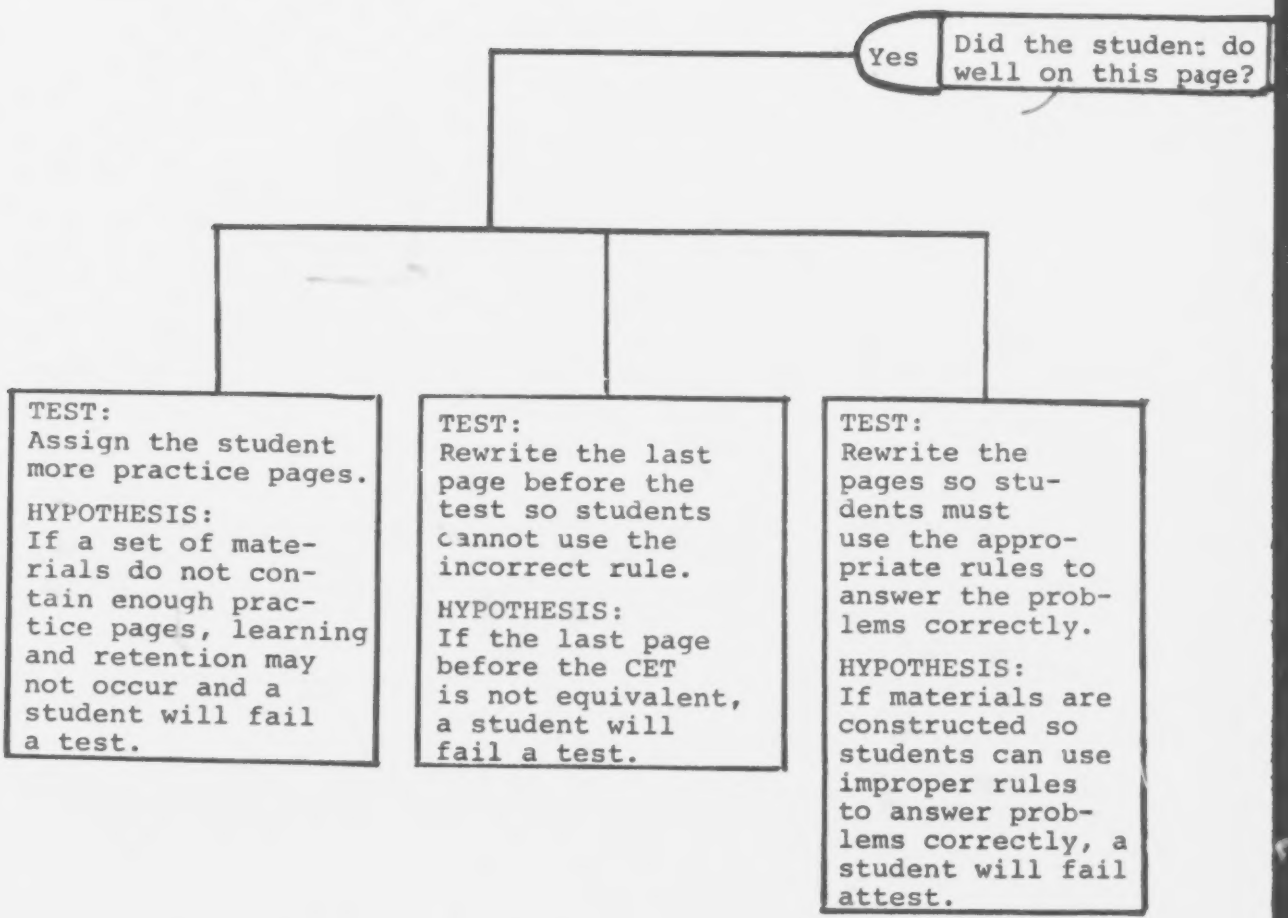

Yes
Did the student do well on this page?
TEST:
Assign the student more practice pages.
HYPOTHESIS:
If a set of materials do not contain enough practice pages, learning and retention may not occur and a student will fail a test.
TEST:
Rewrite the last page before the test so students cannot use the incorrect rule.
HYPOTHESIS:
If the last page before the CET is not equivalent, a student will fail a test.
TEST:
Rewrite the pages so students must use the appropriate rules to answer the problems correctly.
HYPOTHESIS:
If materials are constructed so students can use improper rules to answer problems correctly, a student will fail attest.

FIGURE XII (cont

ANALYSIS OF INSTRUCTIONAL MATERIALS FOR
USING INCORRECT RULE

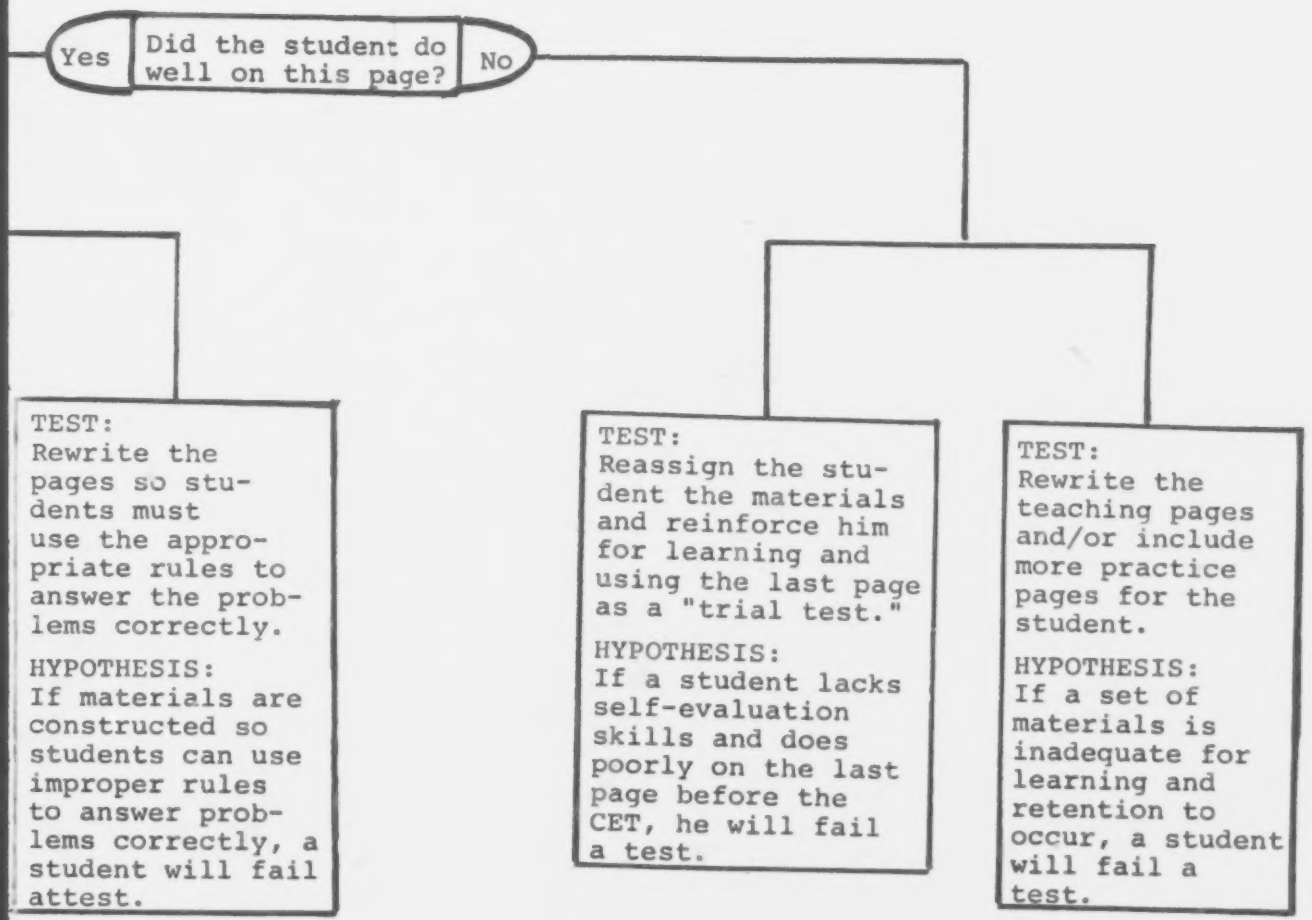
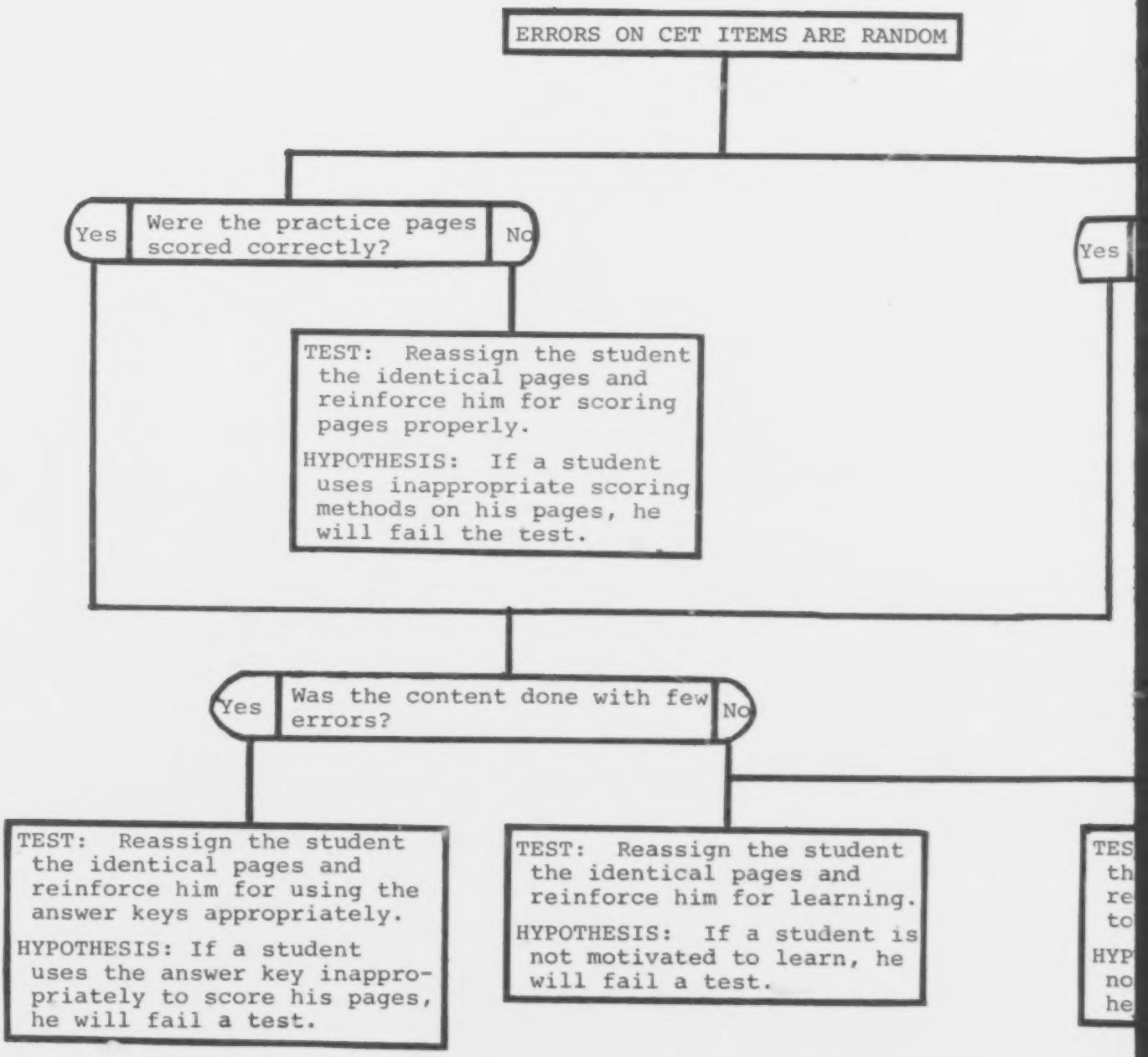

Yes
Did the student do well on this page?
No
TEST:
Rewrite the pages so students must use the appropriate rules to answer the problems correctly.
HYPOTHESIS:
If materials are constructed so students can use improper rules to answer problems correctly, a student will fail attest.
TEST:
Reassign the student the materials and reinforce him for learning and using the last page as a "trial test."
HYPOTHESIS:
If a student lacks self-evaluation skills and does poorly on the last page before the CET, he will fail a test.
TEST:
Rewrite the teaching pages and/or include more practice pages for the student.
HYPOTHESIS:
If a set of materials is inadequate for learning and retention to occur, a student will fail a test.

FIGURE XII (continued)

INSTRUCTIONAL MATERIALS FOR STUDENT USING INCORRECT RULE


ERRORS ON CET ITEMS ARE RANDOM
Yes
Were the practice pages scored correctly?
No
TEST: Reassign the student the identical pages and reinforce him for scoring pages properly.
HYPOTHESIS: If a student uses inappropriate scoring methods on his pages, he will fail the test.
Yes
Yes
Was the content done with few errors?
No
TEST: Reassign the student the identical pages and reinforce him for using the answer keys appropriately.
HYPOTHESIS: If a student uses the answer key inappropriately to score his pages, he will fail a test.
TEST: Reassign the student the identical pages and reinforce him for learning.
HYPOTHESIS: If a student is not motivated to learn, he will fail a test.

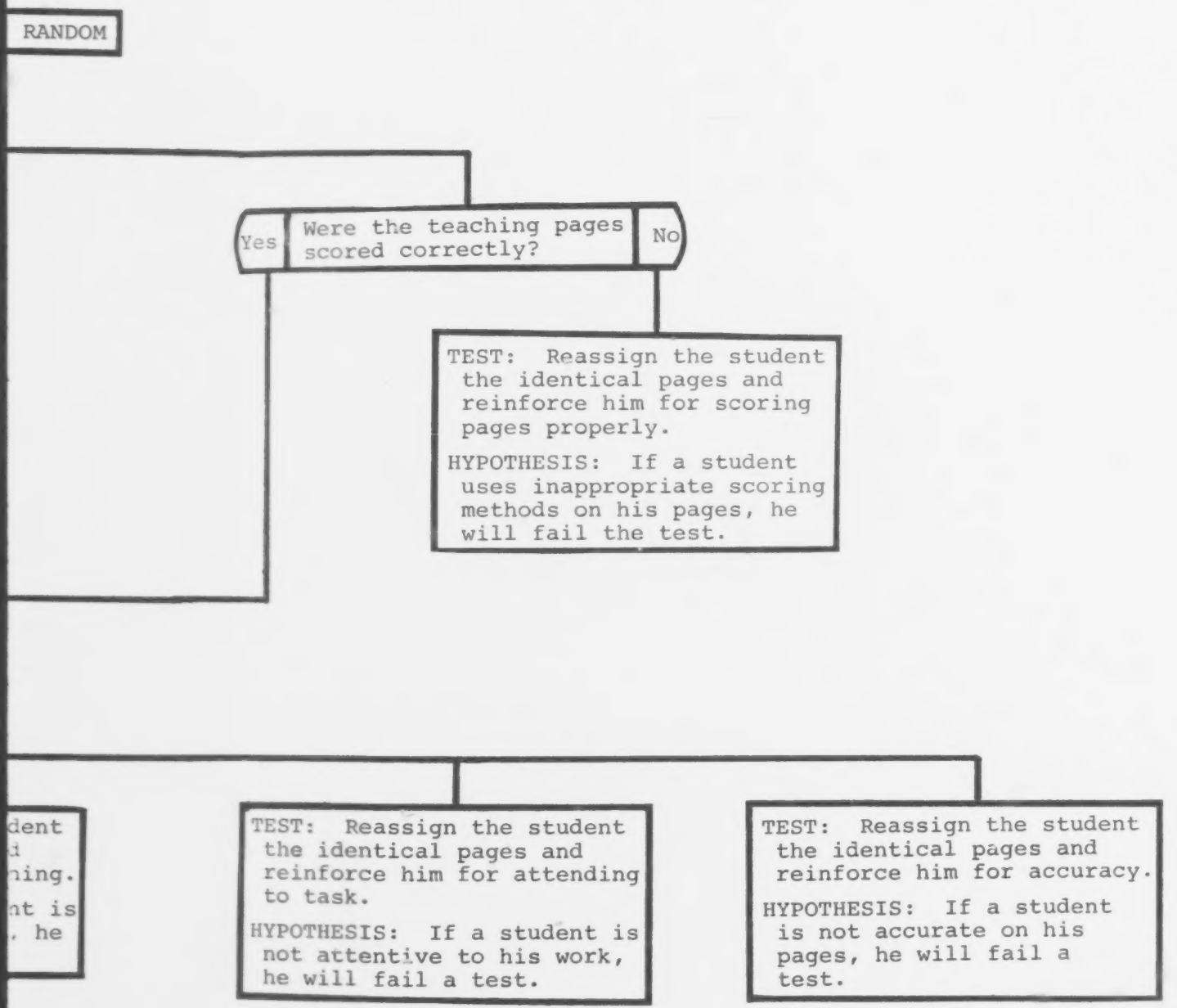
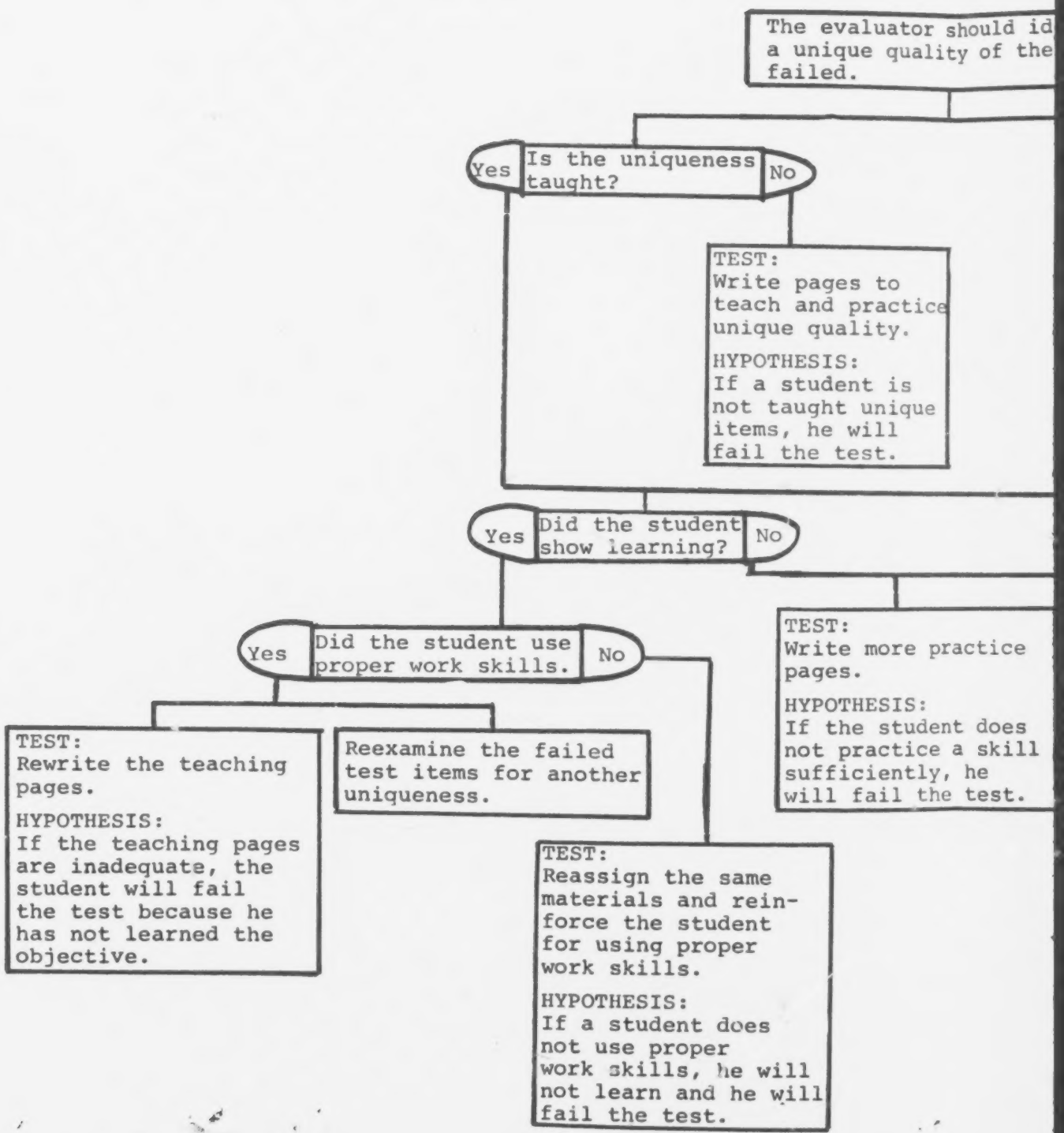

RANDOM
Yes
Were the teaching pages scored correctly?
No
TEST: Reassign the student the identical pages and reinforce him for scoring pages properly.
HYPOTHESIS: If a student uses inappropriate scoring methods on his pages, he will fail the test.
dent
d
ning.
nt is
. he
TEST: Reassign the student the identical pages and reinforce him for attending to task.
HYPOTHESIS: If a student is not attentive to his work, he will fail a test.
TEST: Reassign the student the identical pages and reinforce him for accuracy.
HYPOTHESIS: If a student is not accurate on his pages, he will fail a test.

FIGURE XIII

ANALYSIS OF INSTRUCTIONAL MATERIALS FOR IMPROPER STUDENT BEHAVIOR


The evaluator should id
a unique quality of the
failed.
Yes
Is the uniqueness
taught?
No
TEST:
Write pages to
teach and practice
unique quality.
HYPOTHESIS:
If a student is
not taught unique
items, he will
fail the test.
Yes
Did the student
show learning?
No
TEST:
Write more practice
pages.
HYPOTHESIS:
If the student does
not practice a skill
sufficiently, he
will fail the test.
Yes
Did the student use
proper work skills.
No
TEST:
Rewrite the teaching
pages.
HYPOTHESIS:
If the teaching pages
are inadequate, the
student will fail
the test because he
has not learned the
objective.
Reexamine the failed
test items for another
uniqueness.
TEST:
Reassign the same
materials and rein-
force the student
for using proper
work skills.
HYPOTHESIS:
If a student does
not use proper
work skills, he will
not learn and he will
fail the test.

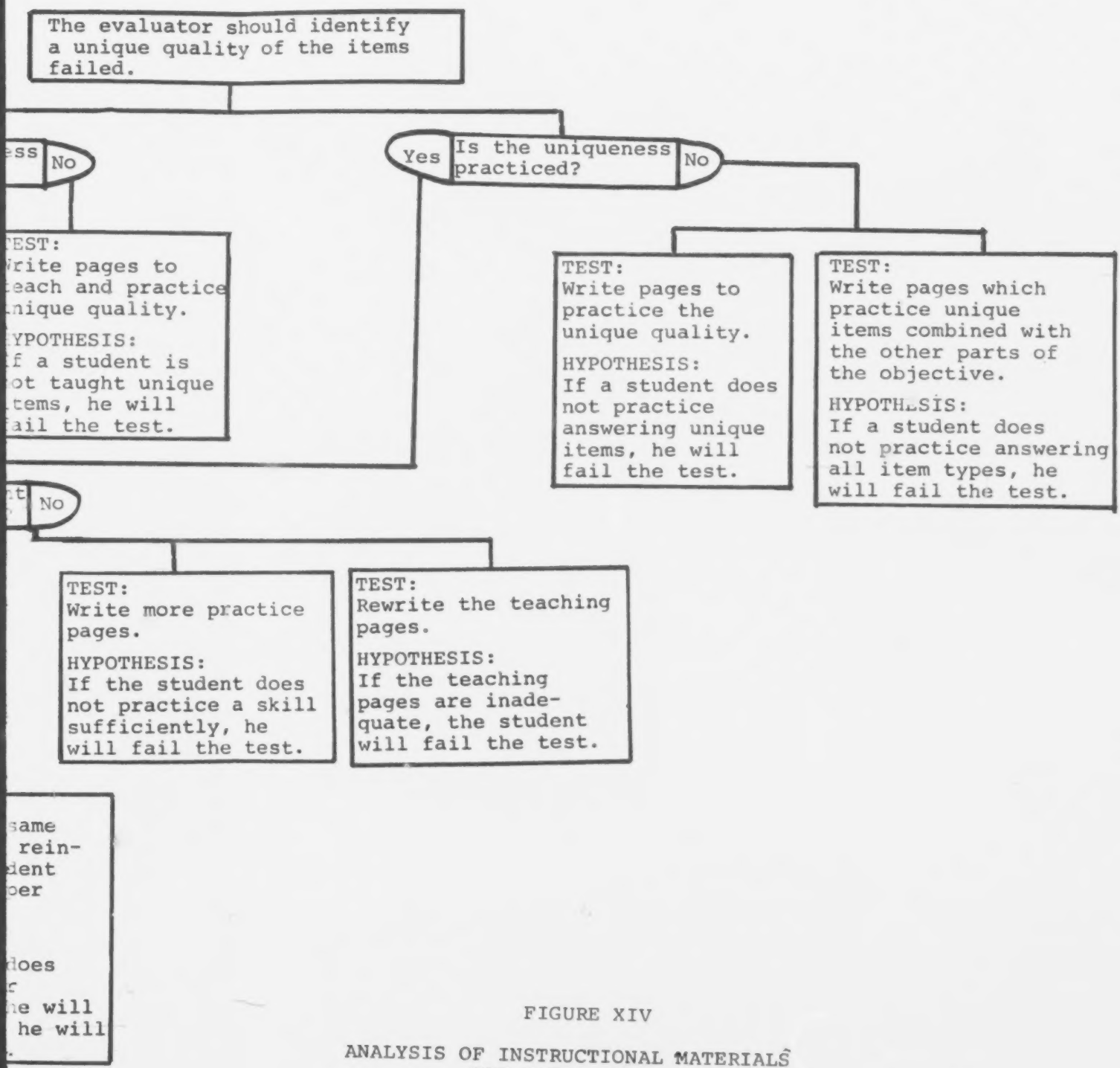

The evaluator should identify a unique quality of the items failed.
ess
No
TEST:
Write pages to teach and practice unique quality.
HYPOTHESIS:
If a student is not taught unique items, he will fail the test.
Yes
Is the uniqueness practiced?
No
TEST:
Write pages to practice the unique quality.
HYPOTHESIS:
If a student does not practice answering unique items, he will fail the test.
TEST:
Write pages which practice unique items combined with the other parts of the objective.
HYPOTHESIS:
If a student does not practice answering all item types, he will fail the test.
No
TEST:
Write more practice pages.
HYPOTHESIS:
If the student does not practice a skill sufficiently, he will fail the test.
TEST:
Rewrite the teaching pages.
HYPOTHESIS:
If the teaching pages are inadequate, the student will fail the test.
same
rein-
ient
ber
does
e will
he will

FIGURE XIV

ANALYSIS OF INSTRUCTIONAL MATERIALS FOR UNIQUE ITEMS

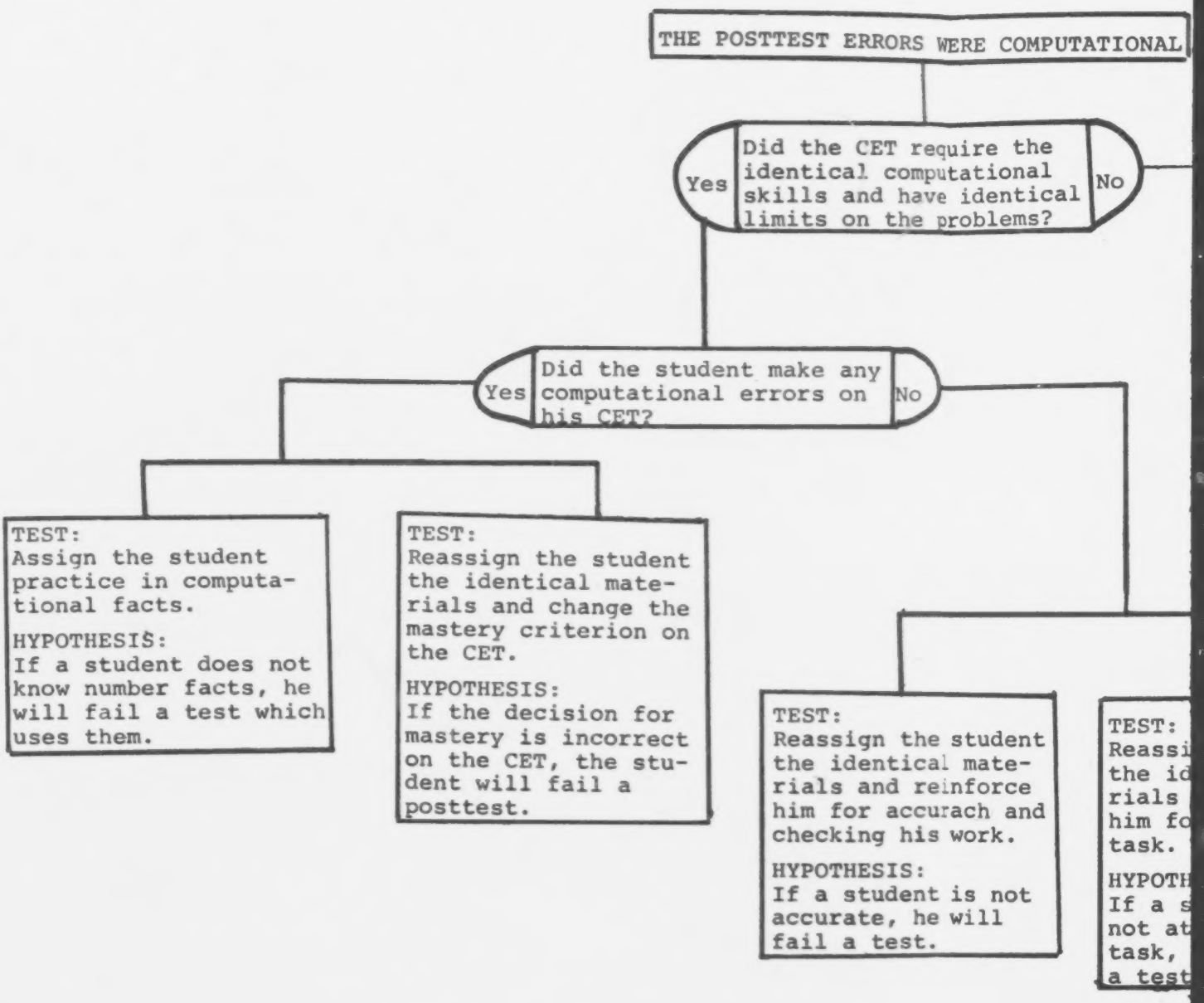

THE POSTTEST ERRORS WERE COMPUTATIONAL
Did the CET require the identical computational skills and have identical limits on the problems?
Yes
No
Did the student make any computational errors on his CET?
Yes
No
TEST:
Assign the student practice in computational facts.
HYPOTHESIS:
If a student does not know number facts, he will fail a test which uses them.
TEST:
Reassign the student the identical materials and change the mastery criterion on the CET.
HYPOTHESIS:
If the decision for mastery is incorrect on the CET, the student will fail a posttest.
TEST:
Reassign the student the identical materials and reinforce him for accurach and checking his work.
HYPOTHESIS:
If a student is not accurate, he will fail a test.
TEST:
Reassi
the id
rials
him fo
task.
HYPOTH
If a s
not at
task,
a test

FIGURE XV

ANALYSIS OF CET FOR COMPUTATIONAL POSTTES

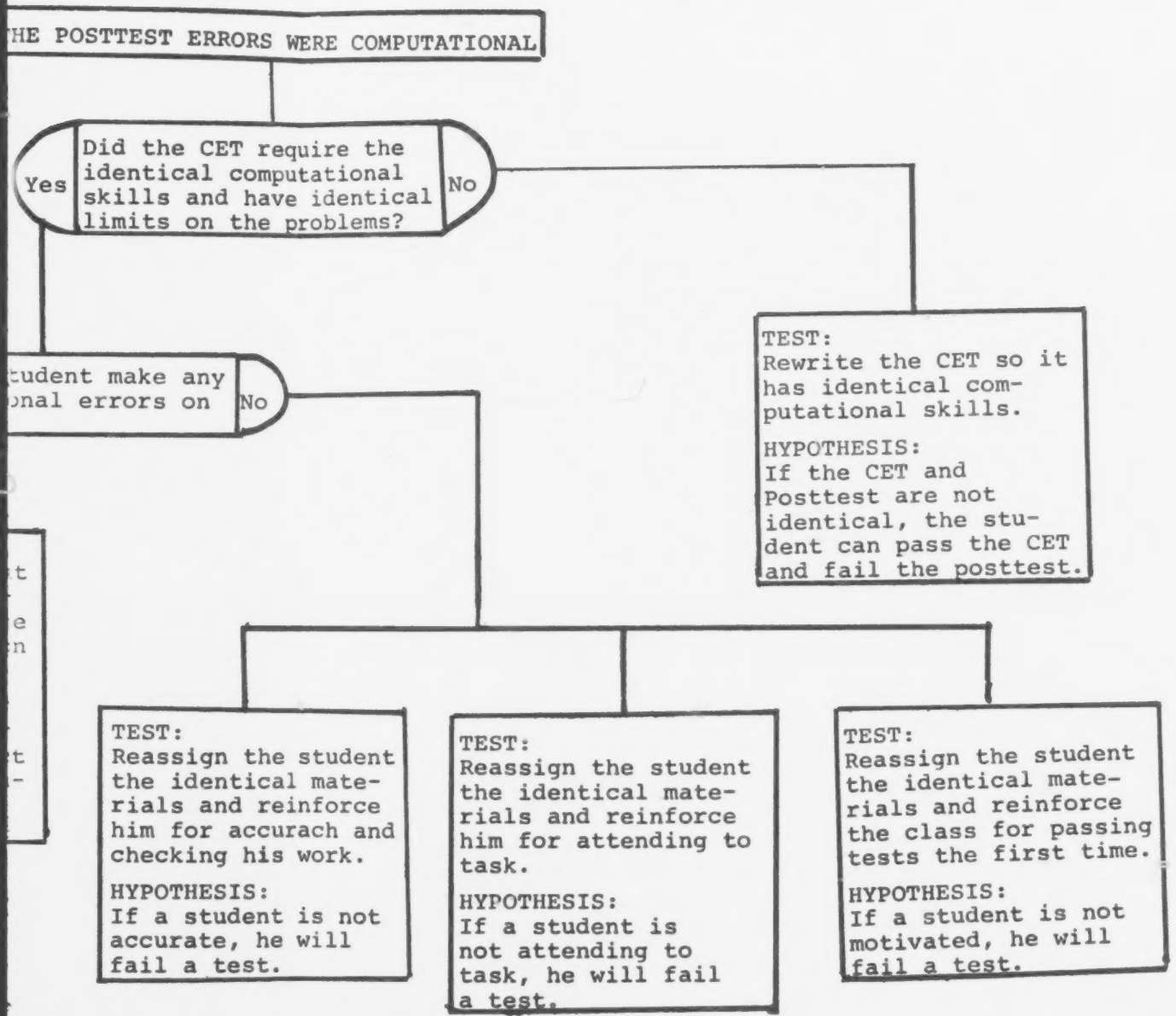

HE POSTTEST ERRORS WERE COMPUTATIONAL
Did the CET require the identical computational skills and have identical limits on the problems?
Yes
No
tudent make any onal errors on
No
TEST:
Rewrite the CET so it has identical computational skills.
HYPOTHESIS:
If the CET and Posttest are not identical, the student can pass the CET and fail the posttest.
TEST:
Reassign the student the identical materials and reinforce him for accurach and checking his work.
HYPOTHESIS:
If a student is not accurate, he will fail a test.
TEST:
Reassign the student the identical materials and reinforce him for attending to task.
HYPOTHESIS:
If a student is not attending to task, he will fail a test.
TEST:
Reassign the student the identical materials and reinforce the class for passing tests the first time.
HYPOTHESIS:
If a student is not motivated, he will fail a test.

FIGURE XV

YSIS OF CET FOR COMPUTATIONAL POSTTEST ERRORS

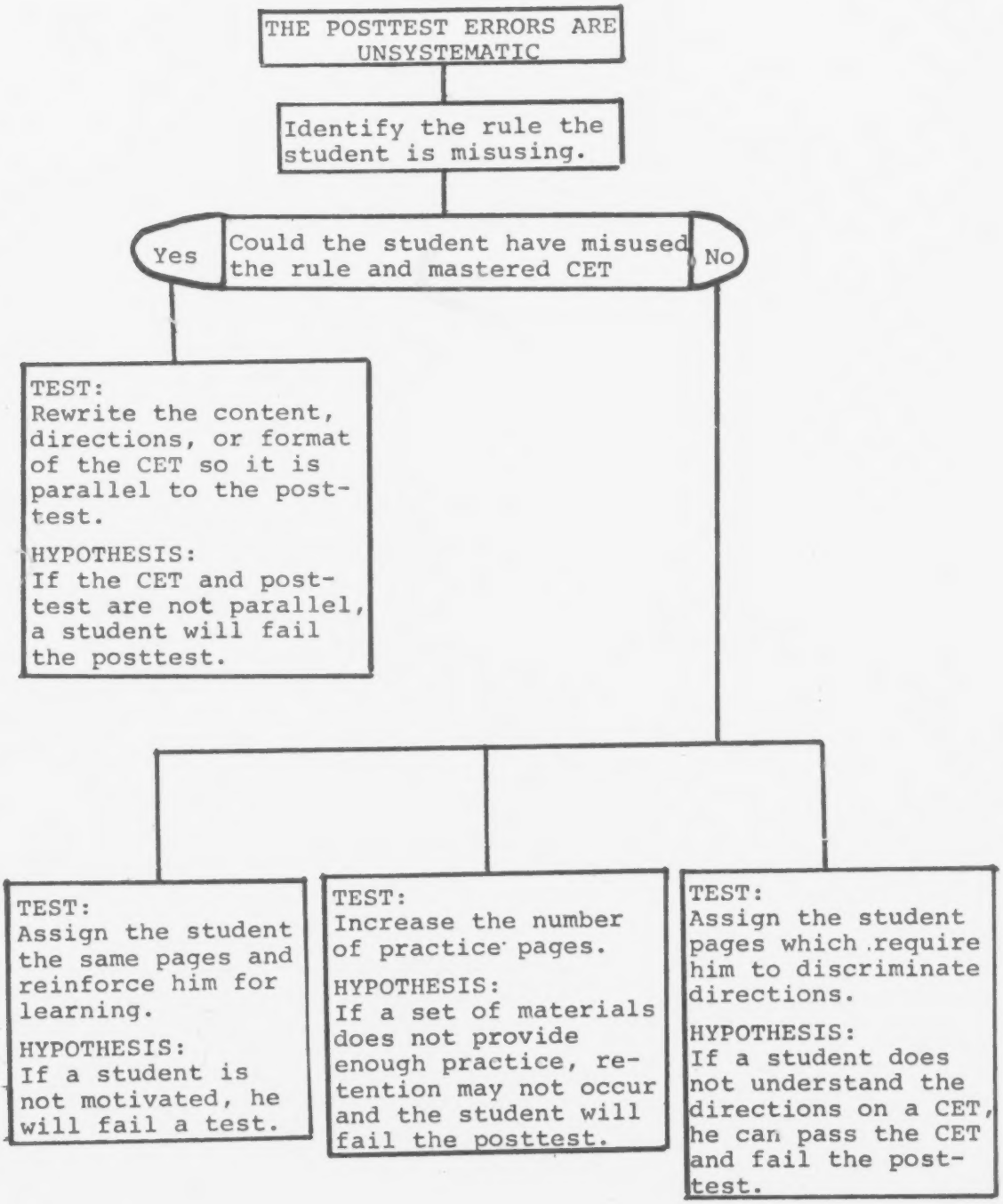

THE POSTTEST ERRORS ARE UNSYSTEMATIC
Identify the rule the student is misusing.
Yes
Could the student have misused the rule and mastered CET
No
TEST:
Rewrite the content, directions, or format of the CET so it is parallel to the post-test.
HYPOTHESIS:
If the CET and post-test are not parallel, a student will fail the posttest.
TEST:
Assign the student the same pages and reinforce him for learning.
HYPOTHESIS:
If a student is not motivated, he will fail a test.
TEST:
Increase the number of practice pages.
HYPOTHESIS:
If a set of materials does not provide enough practice, retention may not occur and the student will fail the posttest.
TEST:
Assign the student pages which require him to discriminate directions.
HYPOTHESIS:
If a student does not understand the directions on a CET, he can pass the CET and fail the post-test.

FIGURE XVI

ANALYSIS OF SYSTEMATIC POSTTEST ERRORS

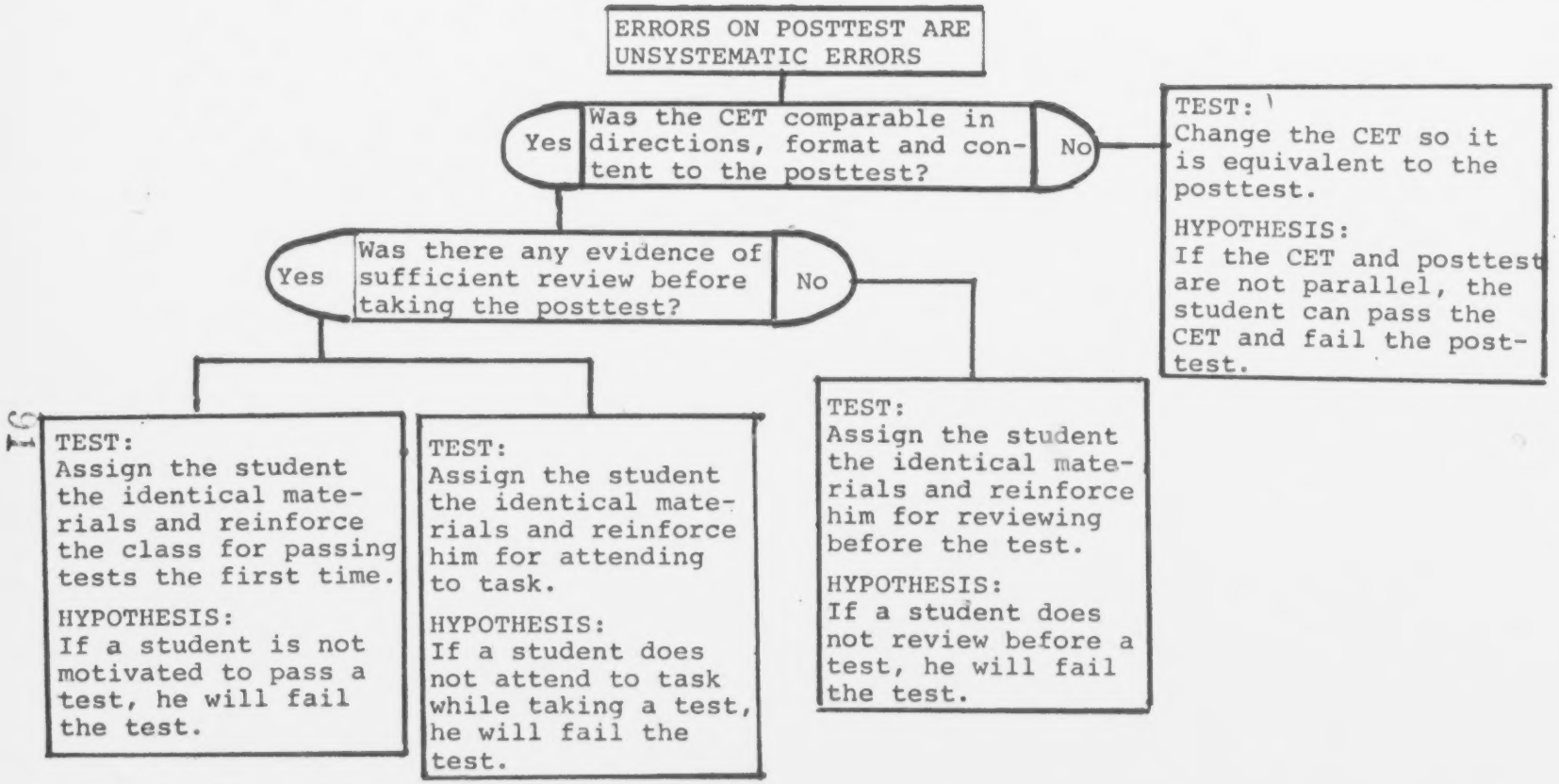

ERRORS ON POSTTEST ARE UNSYSTEMATIC ERRORS
Was the CET comparable in directions, format and content to the posttest?
Yes
No
TEST:
Change the CET so it is equivalent to the posttest.
HYPOTHESIS:
If the CET and posttest are not parallel, the student can pass the CET and fail the posttest.
Was there any evidence of sufficient review before taking the posttest?
Yes
No
TEST:
Assign the student the identical materials and reinforce the class for passing tests the first time.
HYPOTHESIS:
If a student is not motivated to pass a test, he will fail the test.
TEST:
Assign the student the identical materials and reinforce him for attending to task.
HYPOTHESIS:
If a student does not attend to task while taking a test, he will fail the test.
TEST:
Assign the student the identical materials and reinforce him for reviewing before the test.
HYPOTHESIS:
If a student does not review before a test, he will fail the test.

FIGURE XVII

ANALYSIS OF UNSYSTEMATIC POSTTEST ERRORS

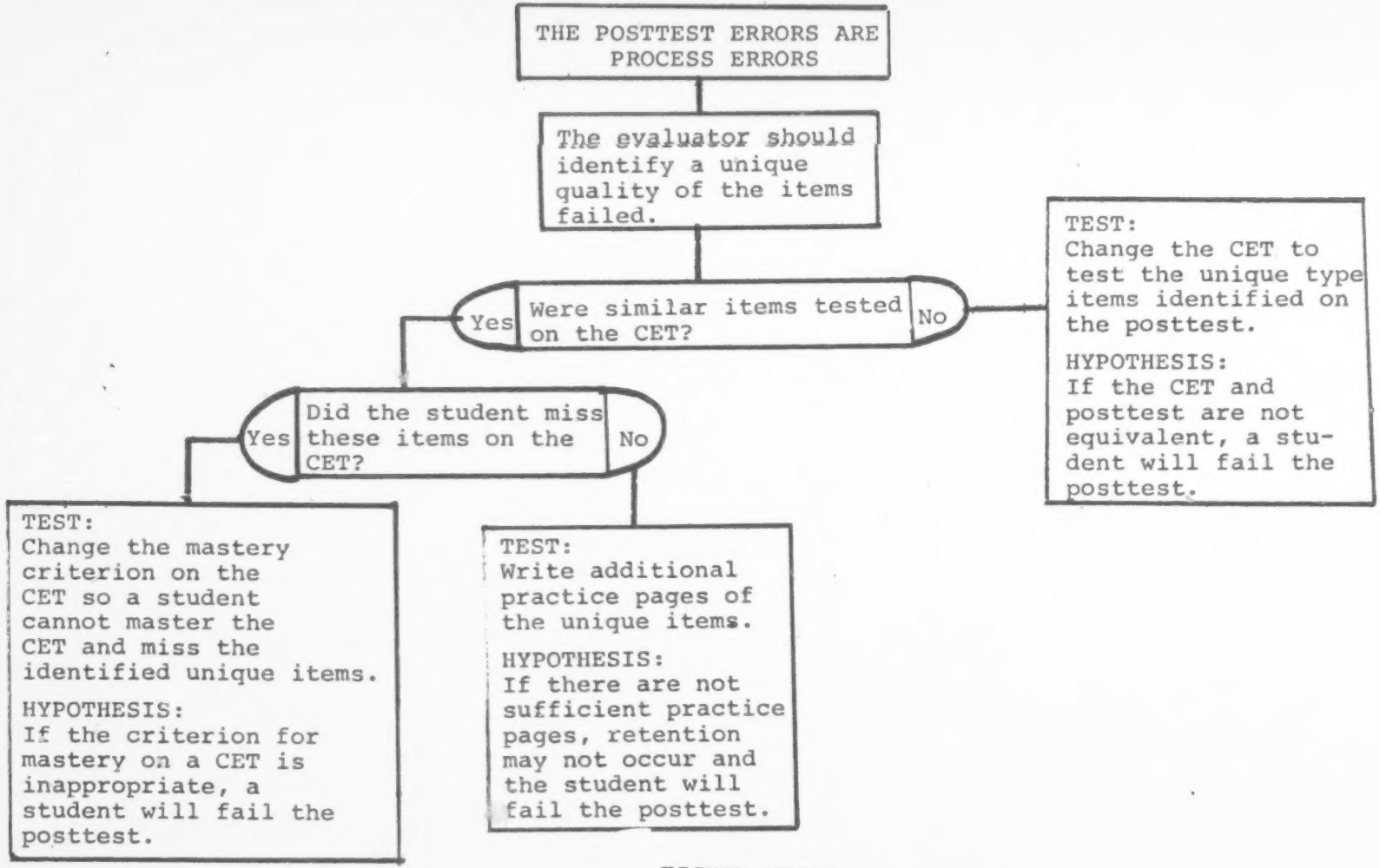

THE POSTTEST ERRORS ARE PROCESS ERRORS
The evaluator should identify a unique quality of the items failed.
Were similar items tested on the CET?
Yes
No
TEST:
Change the CET to test the unique type items identified on the posttest.
HYPOTHESIS:
If the CET and posttest are not equivalent, a student will fail the posttest.
Did the student miss these items on the CET?
Yes
No
TEST:
Change the mastery criterion on the CET so a student cannot master the CET and miss the identified unique items.
HYPOTHESIS:
If the criterion for mastery on a CET is inappropriate, a student will fail the posttest.
TEST:
Write additional practice pages of the unique items.
HYPOTHESIS:
If there are not sufficient practice pages, retention may not occur and the student will fail the posttest.

FIGURE XVIII

ANALYSIS OF PROCESS POSTTEST ERRORS

questions usually result in the evaluator selecting the most appropriate hypothesis to test first.

Although these structures can help the evaluator in selecting a testable hypothesis, they do not identify how the appropriate component of the instructional system should be changed. The evaluator, after examining the items passed and failed on the test and the lesson materials, must decide what the specific deficiency is and how to correct it. For example, knowing that the materials did not teach a skill adequately does not provide information as to how to improve the quality of the teaching pages.

There are several other questions that an evaluator can ask to help him decide how the materials are inadequate. By intensively examining each student's pages, the evaluator can notice where the quality of the student's work began to decrease. If the student started making errors on the pages designed to teach a process, then that teaching page may be poor. If the student made errors on the pages practicing a process, an analysis of the types of errors made on that page can often help the evaluator locate more specifically exactly what the student cannot do correctly.

All decisions made by the evaluator must be tested. If the hypothesis selected to be tested is true, the student should pass the equivalent test; if

the hypothesis is not true, the student should fail the equivalent test. There are always two possible explanations why the hypothesis may prove not to be true: the evaluator has selected the wrong hypothesis or the evaluator has made inadequate revisions to overcome the hypothesis.

Because these procedures require the evaluator to continually test hypotheses and revise materials until the student passes an equivalent test, eventually an appropriate cause of failure is located.

F. The Results of the Formative Evaluation Of IPI Materials

The major emphasis during this study was the development of systematic procedures for the in-context evaluation of instructional materials. There were several indications that the results of using these procedures was the identification and improvement of the IPI materials.

Two major causes of CET test failure were identified: inadequate instructional materials and inappropriate student use of the instructional materials. An analysis of the causes of CET failure attributable to each major cause of failure illustrates the importance of considering all variables as probable hypotheses. Only thirty-five percent of CETs failed the first time

they were taken were caused by inadequate instructional materials. By considering and testing the student's misuse of the materials as a cause of test failure, the evaluator did not have to revise sixty-five percent of the materials associated with test failures.

It is very difficult to report exactly how effective these procedures were in improving instructional materials. For many objectives only a few students used the instructional materials. Therefore, it is difficult to evaluate the effectiveness of one revision used by only one or two students. Gross analysis does indicate that improvements were made in the curriculum during the school year. An analysis of student Posttest performance from the previous school across all grades served to identify fifty-five objectives where less than eighty-five percent of the students passed the objective. This number reflects tests taken in all IPI grades in one school where classroom management procedures were inconsistent; it is quite possible that the figures for test passing are spuriously high. At the end of the school year in which the study was conducted, anaanalysis of student Posttest performance on these same fifty-five objectives was made. Student performance on twenty-seven of the objectives improved to a passing rate the first time of 85% or better, on eighteen of the objectives student performance improved,

on seven of the objectives student performance remained similar, and on the remaining three objectives student performance regressed. This means that student performance was improved on 82% of the objectives during the school year.

CHAPTER V

CONCLUSIONS AND SUMMARY

The purpose of this study was to develop and apply specific evaluation procedures for use with one phase of curriculum development, namely the in-context tryout of instructional materials. These formative evaluation procedures use the method of strong inference to identify inadequate aspects of instructional materials as they are being developed in an on-going classroom.

A. Conclusions

One major problem in evaluating instructional materials during the in-class tryout has been in identifying and controlling many variables which can affect performance on lessons. The procedures developed in this study have been successful in monitoring the total classroom environment by systematically identifying variables which can interfere with pupils learning from lesson materials.

Another major problem in the in-context evaluation of instructional materials has been in selecting a design which can establish cause and effect relationships

between the quality of instructional materials and student performance. The results of this attempt at defining and applying strong inference procedures for evaluating instructional materials appear to be successful and, hence, to suggest an effective design for this phase of curriculum development and evaluation.

B. Summary

The general procedures involved in using strong inference were: (1) defining the specific goals that the instructional materials are designed to achieve, (2) specifying what evidence could be used to identify an instance of pupil failure or non-achievement of such goals, (3) identifying the variables which could account for such failure, (4) generating a list of all probable hypotheses relating such causes to pupil failure, and (5) designing and carrying out crucial experiments to test such hypotheses.

In applying these procedures to the IPI math program, pupil failure on a unit posttest or on a CET was used as the needed specific evidence of lack of achievement. Two sources of information were used to generate probable causes of test failure: an analysis of the student's instructional materials and the observation of the total instructional plan in operation. If the observation of the plan in operation indicated

possible ineffective practices within the classroom environment, the plan was immediately modified. If the analysis of the student's instructional materials indicated several probable causes of test failure, the evaluator designed and carried out experiments to investigate each possible hypothesis.

The steps involved in designing experiments to test each rival hypothesis required the evaluator to pinpoint the specific cause of each test failure. The design of each experiment involved changing one dimension of the instructional system and then retesting the student. If necessary, this process was repeated until an equivalent test was passed. Once the student passed the test, the specific hypothesis being tested was accepted as the probable cause of failure. This identification of a cause of failure led to efficient procedures for immediately correcting either the materials or the plan.

The results presented in the preceding illustrations demonstrate how these procedures can be applied to specific components of the instructional system, namely how one test failure by one student can be used to identify and correct inadequacies within a specific part of the system. The procedures were found to be applicable to other objectives. The result of acculumating all probable causes of test failures for several examples

was a comprehensive generalized inference tree which was usable with other objectives. Once a comprehensive tree was developed from analyzing parts of the system, the evaluator was in a position to use the tree to locate the most probable hypothesis for all tests in the entire system.

Several benefits result from using these procedures to evaluate instructional materials: (1) The procedures appear to be sensitive to errors caused by all components of the instructional system. This sensitivity to all components allows the evaluator to identify and improve simultaneously all inadequacies within an instructional system. (2) The successful use of these procedures is heavily dependent on the evaluator's skill in hypothesizing why a student failed a test and revising materials to improve these inadequacies. Since the evaluator must locate the cause of all failures by performing crucial experiments, his skills are continually being evaluated and improved. (3) All causes of failures must be objectively tested and evaluated. The evaluator must find a hypothesis which, when tested, results in the student passing an equivalent form of the failed test. If the evaluator tests an inappropriate hypothesis, the student should fail the equivalent test. If the student passes the equivalent test, a cause of failure is established and other alternate hypotheses are rejected.

In order to have used these procedures effectively, several conditions were necessary: (1) The complete collaboration of the classroom teacher was an absolute necessity. If the hypothesized cause of failure concerned student study skills or student motivation, the teacher had to be willing to change his classroom behavior in order to test the hypothesis. (2) The on-the-scene presence of the evaluator and lesson writer was also essential. One strength of using strong inference is that the rapid testing of hypotheses is possible. This could only be accomplished if the evaluator was present daily. (3) The evaluator had to be working in a situation where he had the freedom to make daily decisions about revisions. For strong inference to be effective in establishing cause and effect relationships, hypotheses must be continually tested. Once a testable hypothesis is located, either the materials or student behavior are immediately modified and reevaluated.

Several implications concerning curriculum development can be drawn from the results of using these procedures. A close relationship between the lesson writer, formative evaluator, test writer, and objective writer appears essential to curriculum development. In order for a set of instructional materials to be adequate, all components of the instructional system must be focused on identical goals. If the evaluation during the in-context

tryout of materials locates inadequately defined objectives or inadequately written lessons or tests, revisions should be made and tested immediately. This can only be accomplished if the lesson developer and objective developer have a close working relationship since a change in any one component of the system could interfere with the goals of the other components. The members of a curriculum development team should never work independently of each other: each component of the instructional system is dependent on the other components.

The students themselves appear to be the best editors of instructional materials. Their completed materials demonstrated that a student's interpretations of directions and examples may be very different from the curriculum designer's intent. Using the student's completed lesson materials contributed to more understanding about what interferes with a student's learning.

The most obvious implication for curriculum development is that formative evaluation in a classroom is a useful and necessary phase in developing good instructional materials because its results can lead to the immediate improvement of all classroom variables which can interfere with student academic performance.

C. Suggestions for Additional Research

The application of the procedures developed in this study has demonstrated that the formative evaluation

of instructional materials in an on-going classroom is possible and useful. Obviously further studies involving the refinement of these techniques and their use with other types of instructional systems should be considered.

The method of strong inference appears to provide a viable design in the area of curriculum evaluation and development. Although the use of these procedures indicates that they can identify and improve inadequate instructional materials, the specific procedures need additional tryouts. Strong inference, in theory, never establishes direct causal relationships; hypotheses are only accepted until they can be disproven. Additional evaluations of instructional materials are needed to measure the effectiveness of the final product of the procedures. Research concerning how long formative evaluation of materials in the in-context tryout setting should be conducted before a set of instructional materials is accepted and how effective materials evaluated in the in-context setting are in the field testing setting are obvious next steps in refining these procedures.

The procedures described in this study were successful in identifying inadequate paper and pencil self-instructional mathematics materials; their use with other types of materials, such as manipulatives or group lessons, should also be explored. Although different types of materials may require the evaluator to define a different

criterion for locating inadequacies within the instructional materials, to specify different hypotheses concerning causes of failure, and to use different sources of information to locate why the instructional materials are inadequate, the method of strong inference should be equally effective. Studies should be conducted to refine these techniques so that they can be usable with other types of instructional systems.

APPENDIX

APPENDIX A

Classroom Management Procedures

I. Teacher Behavior in the Classroom

A. The teacher walks around the room continuously.

1. Briefly attends to working children
2. Spends no more than one minute with each child
3. Looks at workpages for accuracy
4. Comments on good prescription writing
5. Attends to students who follow rules
6. Reinforces children

B. Upon approaching a student, the teacher will follow the following procedure:

1. Watches the pupil to see if he is working
2. Reinforces working behaviors--if more than one is present, the teacher will reinforce the most complex.
3. Gives prompts to help child, if necessary.
4. Looks at children in direction of travel to see how they are doing, before approaching them.

II. Procedural Rules for the Classroom

1. Begin work as soon as you get your folder.
2. When requesting aid from the teacher, signal with a flag (a folded piece of colored construction

paper) or your hand.

3. Score your own workpages one at a time. You may leave your answer key on your desk during the period.
4. If you have incorrectly answered any problems, go back over your mistakes.
5. When you feel you are ready to take a CET, take your math folder to the aide and your pencil and the CET to the testing area of the room.
6. Check every problem, once you have completed the test.
7. When you have finished your CET, take it to the aide who will give you a scoring key and scoring pen (red).
8. Score your own CET.
9. After your CET is scored, turn it in to the aide.
10. Write your next prescription using the Math Manual.
11. Before taking a posttest, review all skills in the unit.
12. Take your posttest in the testing area. When finished, give it to the aide to correct.
13. While waiting for your posttest to be scored, look over the next unit. While waiting for the pretest to be scored, plan your prescription for the first skill of the next unit.
14. Show all your work. The teacher will ask to see how you arrived at your answers.

Bibliography

Campbell, D. T. From description to experimentation: Interpreting trends as quasi-experiments. In C. W. Harris (Ed.), Problems in measuring change. Wisconsin: University of Wisconsin Press, 1963.

Campbell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally and Company, 1963.

Chitayat, L. An investigation of formative evaluation procedures for use in the in-context tryout of lesson materials and associated instructional procedures. Unpublished masters thesis, University of Pittsburgh, 1970.

Conrad, H. S. The experimental tryout of test materials. In E. F. Lindquist (Ed.), Educational Measurement. Washington, D. C.: American Council on Education, 1951.

Cronbach, L. J. Course improvement through evaluation. Teachers College Record, 1963, 64, 672-83.

Gilbert, T. F. Mathematics: The design of teaching exercises. Journal of Mathematics, 1962.

Hastings, T. J. Curriculum evaluation: The why of outcomes. Journal of Educational Measurement, 1966, 3, 27-32.

Light, J. A., Reynolds, L. J., & Mueller, F. L. Academic performance as a function of teacher attention as a reinforcer. Unpublished.

Lindvall, C. M. & Cox, R. C. with Bolvin, J. O. Evaluation as a tool in curriculum development: The IPI evaluation program. AERA Monograph Series on Curriculum Evaluation, No. 5. Chicago: Rand McNally and Company, 1970.

Lumsdaine, A. A. Educational technology, programmed learning and instructional science. Theories of learning and instruction. Chicago: NSSE, 1964, Part I.

Lumsdaine, A. A. Assessing the effectiveness of educational programs. In Robert Glaser (Ed.), Teaching machines and programmed learning II, DAVI, NEA, 1965.

Markle, S. M. Good frames and bad. New York: John Wiley and Sons, Inc., 1964.

Mueller, F. L., Light, J. A. & Reynolds, L. J. The effects of two styles of tutoring on academic performance. Paper presented at the meeting of the American Educational Research Association, New York, 1971.

Nitko, A. J. Some considerations when tryout out new curricular materials: An antithesis to the theory of chaos. Unpublished draft, University of Pittsburgh, Learning Research and Development, October 27, 1968.

Platt, J. R. Strong inference. Science, 1964, 146, No. 3642, 347-353.

Reynolds, L. J., Light, J. A. & Mueller, F. L. The effects of reinforcing quality or quantity on academic performance. Paper presented at the meeting of the American Educational Research Association, New York, 1971.

Scriven, M. The methodology of evaluation. In Ralph W. Tyler, Robert M. Gagne, and Michael Scriven (Eds.), Perspectives of curriculum evaluation, AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally and Company, 1967.

Sidman, M. Tactics of scientific research. New York: Basic Books, Inc., 1960.

Stake, R. E. Toward a technology for the evaluation of educational programs. In Ralph W. Tyler, Robert M. Gagne, and Michael Scriven (Eds.), Perspectives of curriculum evaluation, AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally and Company, 1967.

Stufflebeam, D. L. The use of experimental design in educational evaluation. Paper presented at the meeting of the American Educational Research Association, Minneapolis, Minnesota, 1970.

Taber, J. I., Glaser, R. H. & Schaefer, H. H. Learning and programmed instruction. Reading, Massachusetts: Addison-Wesley, 1965.